DEEPSURVEY: AGENT-ORIENTED AUTOMATED SURVEY GENERATION WITH ANALYTICAL DEPTH AND CITATION RELIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

As scientific literature grows rapidly and research increasingly involves AI agents, automated survey generation has become a key capability for both agents and human researchers. For such agents, a survey serves as a primary knowledge source of a field prior to research. Since pretrained models already encode broad knowledge of established work, these consumers benefit more from analytical depth than from breadth alone; moreover, unsupported claims, once ingested as knowledge, can propagate into downstream research. However, existing systems tend to overemphasize coverage and presentation; they suffer from limited analytical depth due to reliance on abstracts and isolated paper processing, and from unreliable citations due to imprecise retrieval and post-hoc grounding. We present DeepSurvey, an agentic generation system that addresses both limitations. To enhance depth, DeepSurvey extracts structured keynotes, models cross-paper relationships through clustering and comparative analysis, and integrates a code-agent subsystem to recover implementation-level details. To fortify reliability, it combines citation-graph expansion with hybrid filtering for topic-focused retrieval, enforces evidence-constrained analysis and writing, and deploys multi-granularity agentic refinement to validate citation–claim alignment. Experiments show that DeepSurvey achieves the highest content score (8.34/10) and citation quality (recall and precision gains of 25.3% and 35.2% over the strongest baseline), generalizes more robustly across domains (a CS-to-non-CS score drop of 0.16 vs. 0.28–0.46 for prior systems), and is preferred by domain experts over human-written surveys (83.3% in overall quality, 100% in content depth). Moreover, when a coding agent uses a survey as its only literature source, the agent equipped with DeepSurvey achieves the best performance among human-written and baseline-generated surveys.

1 INTRODUCTION

As scientific literature continues to grow rapidly, AI scientists are increasingly expected to autonomously survey a field, identify open questions, and design experiments (Lu et al., 2024; Yamada et al., 2025), and recent work has demonstrated the feasibility of end-to-end automated survey generation (Wang et al., 2024b). At the same time, the readership of surveys is itself broadening: LLM-based agents increasingly rely on surveys as a primary knowledge source. For these consumers, pretrained models already command broad knowledge of established work, so breadth alone offers relatively marginal value, while analytical depth and reliable attribution matter most. Depth lies in finer-granularity analysis, in implementation-level knowledge that is often omitted or heavily condensed in paper text and recoverable from code. Such depth also lies in connections among papers, from pairwise contrasts to the structure of the whole field. Reliability lies in the correspondence between claims and the evidence they cite. A survey without depth probably only restates what agents already know, whereas a survey without reliability is likely harmful to downstream agents.

However, existing automatic survey generation systems tend to overemphasize coverage and presentation (Wang et al., 2024b), and suffer from three fundamental limitations (Figure 1). *1) Limited survey depth:* Most methods rely solely on abstracts (Wang et al., 2024b; Yan et al., 2025) or domain-specific templates (Liang et al., 2025) that fail to generalize across diverse fields. Fur-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

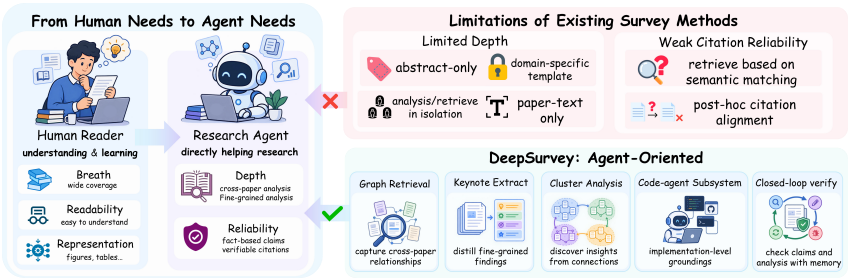


Figure 1: Survey consumers from human to agent. Research agents require analytical depth and verifiable reliability more than breadth and presentation. DeepSurvey addresses the corresponding limitations of existing methods in depth and citation reliability through graph-based retrieval, keynote extraction, cluster analysis, a code-agent subsystem, and closed-loop verification.

thermore, they typically process papers in isolation without fully modeling cross-paper relationships (Go et al., 2025; Chao et al., 2025). Crucially, by overlooking non-textual artifacts such as code repositories (Nguye et al., 2025), these systems fail to capture key technological details and implementation-level insights that are often omitted or heavily condensed in the main text. *2) Weak citation reliability:* On the retrieval front, dense semantic matching often introduces tangentially related papers that dilute the survey content (Wu et al., 2025). On the grounding front, citations are typically attached via post-hoc alignment rather than derived directly from cited evidence, leaving substantial claims without verifiable support (Wang et al., 2024b). *3) Unconvincing evaluation.* Existing evaluations are narrowly restricted to computer science (Shi et al., 2025; Su et al., 2025; Nguye et al., 2025), adopt outdated metrics (Bao et al., 2025) or metrics with weak discriminative power (Wang et al., 2024b), and compare against only a handful of baselines (Wu et al., 2025; Zhou et al., 2025; Liu et al., 2025b), making it difficult to reliably assess and differentiate methods.

To address these limitations, we propose DeepSurvey, an agent-oriented system optimized along three core axes: *1) Enhancing survey depth:* To achieve scalable, domain-agnostic analysis, **DeepSurvey** extracts structured, informative keynotes from full-text literature (§2.2). Moving beyond isolated paper processing, we explicitly model cross-paper relationships by clustering the literature and conducting multi-level comparative analysis (e.g., via structural relation graphs and joint question-answering reasoning). Furthermore, we integrate a dedicated code-agent subsystem to cross-reference official repositories; this allows DeepSurvey to retrieve and complement critical technological dimensions that are missing from pure text, ensuring a more comprehensive and holistic literature synthesis (§2.3). *2) Fortifying citation reliability:* We combine citation-graph expansion with LLM-based re-ranking to maintain strict topical consistency and filter out irrelevant papers (§2.1). Instead of relying on vulnerable post-hoc alignment, we enforce strict source attribution throughout the entire generation lifecycle (§2.3): directly anchoring citations during multi-stage outline planning (§2.4), and deploying an agentic refinement loop to rigorously validate citation-claim pairs (§2.5). *3) Strengthening evaluation:* we construct a cross-domain evaluation protocol grounded in existing influential benchmarks, redesign the evaluation protocol with more discriminative and agent-oriented criteria, and include a broader set of competitive baselines to enable more rigorous and informative comparisons (§3.1).

Extensive experiments demonstrate that DeepSurvey achieves an overall content score of 8.341/10 with a depth score of 7.897/10, substantially surpassing prior methods, and reaches 0.630 recall and 0.569 precision in citation quality, outperforming the strongest baseline by 25.3% and 35.2% (§3.2). DeepSurvey also generalizes more robustly across domains, exhibiting a much smaller performance drop in non-CS fields than existing methods (0.16 vs. 0.28–0.46) (§3.2). Notably, in downstream validation, a coding agent working from a DeepSurvey survey autonomously outperforms the best published result on PINNacle (relative L_2 error 1.53×10^{-2} vs. 3.81×10^{-2}), a result no other literature condition reaches (§3.3). In pairwise human evaluation, domain experts prefer DeepSurvey over human-written surveys, especially in overall quality (83.3% win rate) and content depth (100% win rate) (§3.4). Our evaluation protocol is also validated through meta-evaluation: the LLM judge achieves a low coefficient of variation of 0.244%, and its Cohen's κ with human experts (0.6463) approaches human–human agreement (0.6958) (§3.6).

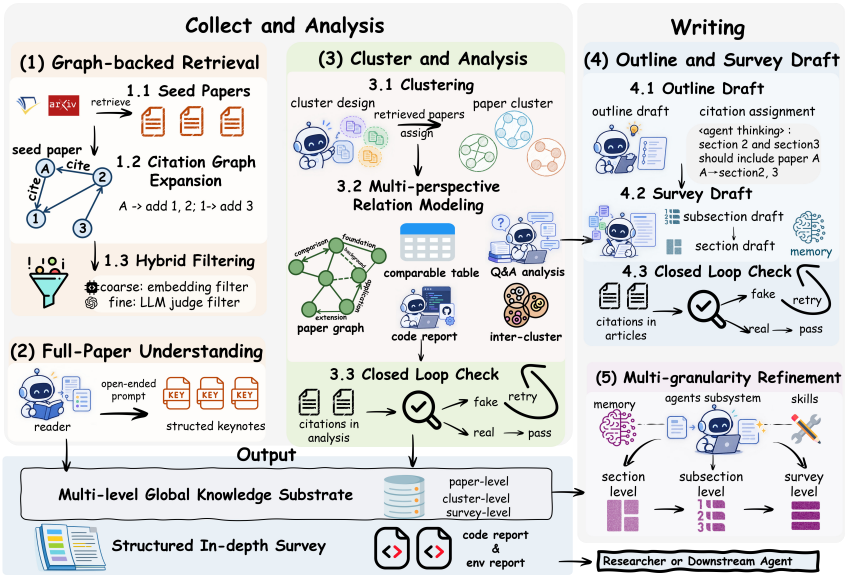


Figure 2: Overview of DeepSurvey. Stages 1–3 build the analysis substrate via graph-based retrieval, keynote extraction, clustering with multi-perspective relation modeling and code-agent subsystem. Stage 4 generates the survey. Stage 5 applies multi-granularity agentic refinement.

Overall, DeepSurvey provides an end-to-end system for generating in-depth and reliable surveys across domains, designed to support literature synthesis especially for AI scientists.

Our main contributions are as follows:

- We propose **DeepSurvey**, the first agent-oriented survey generation system that prioritizes the *analytical depth* agents need over breadth, building a reusable analysis substrate via keynote extraction, cross-paper relation modeling, and **a code-agent subsystem**.
- We design closed-loop **citation reliability** mechanisms across the entire generation pipeline, including graph-based retrieval, evidence-constrained writing, and multi-granularity agentic refinement.
- Extensive experiments show that DeepSurvey achieves the best content and citation quality across domains and human expert preference over human-written surveys. A coding agent working from a DeepSurvey survey also outperforms the best published result on PINNacle.

2 DeepSurvey

In this section, we propose DeepSurvey, an agentic system that produces an agent-oriented survey together with a structured, reusable analysis knowledge substrate (Figure 2): Stages 1–3 build the analysis substrate through graph-based retrieval, full-text keynotes, code-agent analyses and clustering with multi-perspective relation analysis; Stage 4 generates the survey from it with outline-driven planning and citation assignment; Stage 5 applies multi-granularity agentic refinement. Engineering mechanisms for robustness are in Appendix A.3.

2.1 Graph-based Literature Retrieval

Given a research topic, DeepSurvey retrieves seed papers from Semantic Scholar (Wade, 2022; Lo et al., 2020; Xiong et al., 2017), providing broader coverage than static local databases. A judge agent filters initial candidates by topical alignment to reduce seed-stage drift. Starting from the seed set, DeepSurvey expands papers through the edges in citation graph in both directions, capturing inter-paper relations that flat semantic search misses. A two-stage filtering strategy controls expansion noise: a coarse semantic filter (encode abstracts and titles) preserves high-recall candidates, while a fine-grained LLM re-ranking removes topically irrelevant papers.

2.2 Full-text Understanding

For each paper, DeepSurvey reads the full text and generates a structured **keynote** covering contributions, methods, experiments, assumptions, and limitations. Since key technological details and implementation-level insights are often omitted or heavily condensed in abstracts, full-text reading is essential for capturing the fine-granularity information needed for meaningful cross-paper comparison (Cohan & Goharian, 2015). The open-ended structure avoids fixed templates, enabling domain-adaptive representation that supports cross-domain summarization without sacrificing specificity.

2.3 Clustering and Relation Analysis

After keynotes generation, DeepSurvey designs cluster themes based on the complete keynote collection and assigns papers to clusters through a clustering agent and a partitioning agent. Papers spanning multiple cluster topics may be assigned to multiple clusters. A verification-retry loop corrects missing or hallucinated assignments.

Within each cluster, three complementary perspectives are used to model cross-paper analysis: **relation graphs** model typed citation relations (e.g., foundation, extension, substitution) for technical lineage, **comparison tables** compare papers along key dimensions, and **guided Q&A** extracts cross-paper patterns. **Inter-cluster** analysis is also performed to capture cross-cluster differences. A **code-agent subsystem** (Figure 3) explores the associated repositories to recover implementation-level details overlooked by purely textual analysis: a planning agent orchestrates specialized skills to distill repository pseudocode, which is synthesized into repository-level reports and further integrated into a topic-level report; these serve as implementation-level evidence for the writing and refinement agents, indexed in the knowledge substrate (details in Appendix A.1.

Analyses require explicit source paper attribution, with monitoring to detect and correct hallucinated sources. All outputs are stored in the global knowledge substrate (details in Appendix A.2). Cases are in Appendix F.

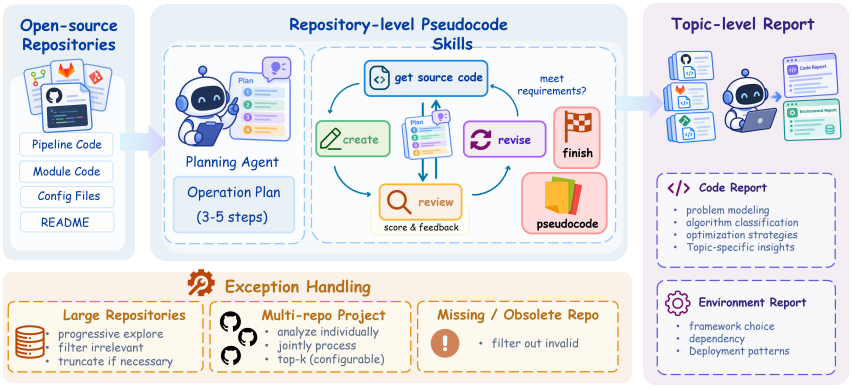


Figure 3: The code-agent subsystem. A centralized planning agent orchestrates specialized skills to iteratively explore open-source repositories and distill repository-level pseudocode with exception handling. The pseudocode is synthesized into topic-level reports.

2.4 Outline Draft and Scoped Writing

DeepSurvey first generates a hierarchical outline from the clustering results with structured evidence in the analysis knowledge substrate, and iteratively refines it against retrieved paper keynotes. The outline specifies the content scope and constraints of each section and subsection.

Following outline generation, papers are *explicitly assigned* to appropriate sections based on keynotes and cluster analyses in knowledge substrate. This citation-anchoring process creates a localized evidence constraint: each writing unit depends only on its assigned paper subset. This mechanism can improve the model's overall performance and reduce hallucination risk by limiting the context (Liu et al., 2023a; Du et al., 2025; Liu et al., 2025a).

Writing proceeds bottom-up. **Subsection drafts** receive the outline, subsection description, relevant cluster-based analyses, and keynotes for the assigned papers from knowledge substrate. **Section drafts** aggregate subsection content with overview-level analysis to form higher-level narratives.

After each paragraph is generated, a **citation verification** mechanism checks for consistency and authenticity against the assigned paper set. Detected citation errors trigger retry generation with error memory, ensuring strict alignment between the generated text and the evidence collection. Compared with previous methods that simply hard-match hallucinated citations against a database, our approach better aligns claims with citations.

2.5 Multi-granularity Refinement

After initial draft generation, DeepSurvey applies multi-granularity refinement at section, subsection, and survey levels to improve content quality and validate citation–claim pairs. A **centralized planning agent** inspects the draft, outline, revision history, and relevant evidence, then coordinates specialized skills: readers retrieve keynotes and source passages; the reviewer assesses quality and provides suggestions; and the reviser applies edits. The planning agent iteratively invokes these skills until the planner considers the draft satisfactory or a maximum round count is reached.

Two mechanisms improve refinement quality. First, **global memory** preserves reviewer feedback, score changes, and revision states across rounds. Second, **multi-granularity optimization** performs section-level refinement for organization and logic, subsection-level refinement for local analysis, readability and citation–claim validation, and survey-level refinement for coherence. We also include a simplified skill-loop fallback for weaker backbone models.

3 Experiments

Reliable evaluation of survey generation requires benchmarks that are both diverse and discriminative. However, existing protocols remain limited: they are either restricted to computer science (Yan et al., 2025) or lack sufficient discriminative power (Table 5, details in Appendix B.1), they often compare against only a narrow set of baselines (Liang et al., 2025), and they rarely assess whether a survey can directly support downstream agents in conducting research. These limitations motivate a more rigorous, multi-dimensional, cross-domain and agent-oriented evaluation setup.

3.1 Experiment Setup

To address these limitations, we construct a multi-domain evaluation setup by combining the complementary strengths of prior benchmark efforts. For topic selection, we draw a high-quality subset from SurveyBench (Yan et al., 2025) and AutoSurvey (Wang et al., 2024b) in computer science, focusing on well-defined topics in LLMs and machine learning. Moreover, as automated research expands to a growing range of disciplines, research agents will increasingly need surveys well beyond computer science; we therefore further select a subset of SurveyGen (Bao et al., 2025) covering high-interest topics in Biology, Medicine, Psychology, etc. (details in Appendix D.1). Note that we use the MiMo-series model (MiMo-V2.5) [1] as the evaluation model (reasons in Appendix B.2); DeepSurvey remains the top method under judges from other model families as well, confirming that the results are not an artifact of judge–generator family affinity (Appendix D.9).

3.1.1 Evaluation Metrics

We evaluate generated surveys along two dimensions: content quality measures whether generated surveys achieve sufficient depth and research value, while citation quality assesses the reliability and accuracy of references (detailed metrics are presented in Appendix D.2).

Content Quality is assessed using an LLM rubric judge on a 1–10 scale. Building on AutoSurvey's influential framework (Wang et al., 2024b) and Agentic AutoSurvey's discriminative standard (Liu et al., 2025b), we further adapt the protocol to the needs of agent consumers (detailed analysis in Appendix D.3). We evaluate three aggregated dimensions: **Core Quality** (synthesis, organization,

[1] https://mimo.xiaomi.com/mimo-v2-5

Table 1: Performance comparison of different survey generation methods on overall, fine-grained content metrics, and citation quality (precision and recall). The best and second-best results are highlighted in purple and light purple, respectively.

Method	Total Score	Content Quality														Citation Quality	
		Core Quality					Writing Quality				Content Depth						
		Total	Synth.	Org.	Comp.	Rel.	Total	Read.	Rigor	Clarity	Total	Crit.	Novel.	Spec.	Future	Precision	Recall
AutoSurvey	7.958	8.646	7.983	8.608	8.718	9.276	8.153	7.606	8.500	8.353	7.171	7.049	6.847	7.991	6.798	0.342	0.440
SurveyX	6.993	8.187	7.283	8.133	8.000	**9.333**	7.067	7.053	6.313	7.833	5.763	5.733	5.133	5.983	6.200	0.125	0.194
SurveyForge	8.071	8.616	7.975	8.700	8.690	9.100	8.153	**7.793**	8.283	**8.382**	7.485	7.350	7.290	7.895	7.405	0.421	0.503
LiRA	5.687	5.049	7.850	8.425	2.565	1.355	6.805	7.695	5.155	7.565	5.766	5.775	5.045	6.095	6.150	0.227	0.442
DeepSurvey	**8.341**	**8.865**	**8.334**	**8.935**	**8.889**	9.300	**8.183**	7.711	**8.506**	8.333	**7.897**	**7.389**	**7.866**	**8.194**	**8.139**	**0.569**	**0.630**

comprehensiveness, relevance), **Writing Quality** (readability, academic rigor, clarity and coherence), and **Content Depth** (critical analysis, novelty and insights, specificity, future directions). The final score is computed as 0.4 $\times$ Core + 0.2 $\times$ Writing + 0.4 $\times$ Depth. Further explanation on the formula and the dimensions is in Appendix D.2.

Citation Quality follows AutoSurvey (Wang et al., 2024b) and is measured by Citation Recall (whether claims are supported by their citations) and Citation Precision (whether cited papers are truly relevant to the claims); the NLI judgments are rendered by the same judge model (MiMo-V2.5). Details are presented in Appendix D.2.

3.1.2 Baselines

We compare DeepSurvey against four representative automated survey generation systems spanning diverse methodological approaches. **AutoSurvey** (Wang et al., 2024b) pioneered the paradigm of semantic retrieval with parallel generation. **SurveyX** (Liang et al., 2025) leverages AttributeTree for template-based literature understanding. **SurveyForge** (Yan et al., 2025) introduces SANA for high-quality retrieval. **LiRA** (Go et al., 2025) employs a multi-agent workflow with specialized roles[2]. All baselines use the same generation backbone as DeepSurvey (MiMo-V2.5). Complete parameter settings are presented in Appendix D.4 and the specific topics for evaluation are presented in Appendix D.1.

3.2 Main Results

Content and Citation Quality. The results are shown in Table 1. For **content quality**, DeepSurvey achieves the highest overall content score (8.341), with Content Depth at 7.897, which is 0.412 above the second-highest baseline. For **citation quality**, DeepSurvey achieves the best scores, with 0.630 recall and 0.569 precision, outperforming the strongest baseline SurveyForge by 25.3% and 35.2%, respectively, highlighting the effect of the systematic reliability mechanism. Overall, the DeepSurvey system improves both survey depth and citation reliability. A **case study** versus the most influential baseline AutoSurvey (Figure 12) further confirms DeepSurvey's superiority: it stays topic focused and provides human-expert-aligned insights, while AutoSurvey shows topic drift and superficial analysis (details in Appendix G.1).

Meta-Analysis. All results are averaged over repeated runs. The evaluation proves stable: aggregated CVs are at most 6.3% across methods, and fine-grained sub-dimensions mostly stay below 10% (Appendix D.5). DeepSurvey itself is highly stable (aggregated CVs of 0.8%–2.2%), with its gains over all baselines reaching large effect sizes (Hedges' $g \geq 0.8$) on Total Score, Content Depth, and citation metrics.

Cross-domain Evaluation. Figure 4 separates fields in Table 1: **DeepSurvey** is consistently the strongest method across CS and non-CS domain settings, surpassing **AutoSurvey** and **SurveyForge**, the two strongest prior baselines, and maintains a much smaller drop when shifting to the non-CS domain than all other methods (0.16 vs. 0.28–0.46), confirming substantially stronger domain generalization.

[2]LiRA does not include a retrieval module so we use AutoSurvey's retrieval module as a substitute.

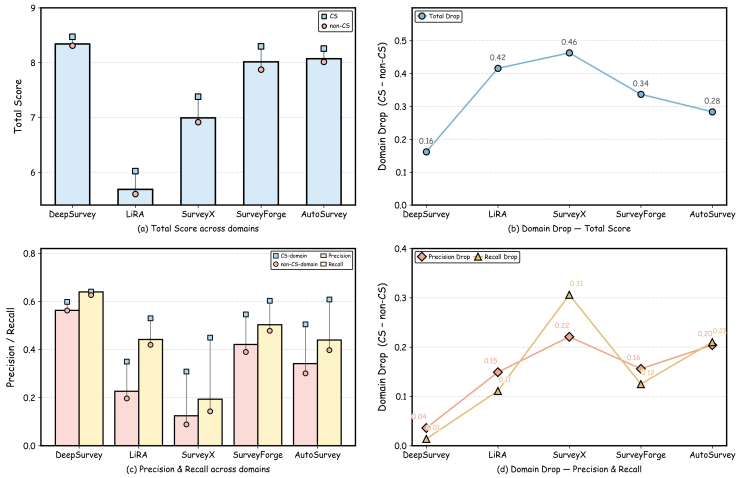


Figure 4: Cross-domain generalization across CS and non-CS domains. The left panels (a, c) compare overall, CS, and non-CS scores for each method; the right panels (b, d) show the absolute total-score gap between CS and non-CS domains.

3.3 Downstream Validation with a Coding Research Agent

To examine whether a DeepSurvey survey improves the research capability of downstream agents, we task an identical coding agent (codex) with optimizing vanilla PINNs on PINNacle (Hao et al., 2023), whose published optimization methods are stripped, under four literature conditions (none, human-written survey, SurveyForge, and DeepSurvey) that serve as its only knowledge source, within a 480-minute budget. Each condition is run three times, and the best run is reported (details in Appendix C). The DeepSurvey-condition transcript shows its mechanism-first survey, which maps each method to the targeted failure mode and its applicability conditions, translated directly into design decisions: the agent measures a per-axis spectral fingerprint from the PDE operator, completes a hard-constraint output transform guided by the survey's analysis, and, finding no dedicated techniques for hyperbolic problems, autonomously invents a symmetric d'Alembert ansatz that yields the Wave-1D-C breakthrough. The delivered method attains an overall relative L_2 error of 1.53×10^{-2} (Table 2), the best of all conditions: $2.5\times$ below the per-case-best published PINNacle reference and the highest novelty score judged by GPT-6 Astra, whereas all other conditions remain above the published reference. This contrast suggests that the implementation-level understanding conveyed by a DeepSurvey survey, rather than the enumeration of methods, is what transfers to downstream agents.

3.4 Human Expert Evaluation

To verify that DeepSurvey is also applicable to human researchers, we further conduct blinded pairwise evaluations where domain experts compare DeepSurvey against human-written surveys on the same topic set, without knowing each response's origin (detailed settings in Appendix D.7). Figure 5 summarizes the results: DeepSurvey is preferred in overall quality (winning rate 83.3%), with the strongest advantage (winning rate 100%) in Content Depth. Writing Quality is comparatively weaker: DeepSurvey wins 50.0% annotations, and win 54.5% once the 2 ties are excluded, which is consistent with the automatic evaluation results.

3.5 Ablation Studies

We evaluate the contribution of each core module by removing it from the full system: **w/o collector** replaces graph-based retrieval with static database lookup; **w/o keynote** uses only abstracts instead of full-text keynotes; **w/o analyzer** skips clustering and relation modeling; **w/o writer** generates directly from retrieval results without outline-driven writing or citation checks; **w/o refinement** disables iterative agentic optimization. **w/o agentic** replaces the central-planner agentic refinement system with a skill-loop (iteratively calling review and revise). **w/o code** skips code-agent subsystem's

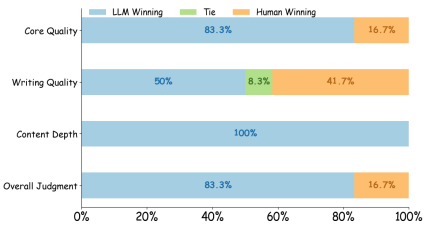


Figure 5: Comparison between DeepSurvey generated and human written surveys on the same topic set judged by human experts.

Condition	Burgers-1D-C	Heat-2D-CG	Heat-2D-MS	Wave-1D-C	Overall	SOTA	Novelty
PINN-base (pilot)	2.10e-02	2.58e-02	1.53e-01	4.86e-01	1.71e-01	–	–
PINNacle SOTA	1.33e-02	2.39e-02	1.73e-02	9.79e-02	3.81e-02	–	–
none	**1.33e-02**	**1.78e-02**	1.85e-02	4.68e-01	1.29e-01	1	5
human	2.20e-02	2.59e-02	4.80e-02	1.63e-01	6.48e-02	0	4
surveyforge	1.88e-02	2.60e-02	9.73e-02	5.10e-01	1.63e-01	0	5
deepsurvey	1.36e-02	2.64e-02	**1.18e-02**	**9.54e-03**	**1.53e-02**	2	7

Table 2: Downstream validation on PINNacle: relative L_2 error (lower is better). "SOTA" indicates number of results that surpass the best published PINNacle result. "Novelty" is scored by GPT-6 Astra (details in Appendix C).

Method	Total	Core	Writing	Depth
w/o collector	6.031	5.082	5.889	7.052
w/o keynote	7.957	8.634	8.081	7.218
w/o analyzer	7.933	8.514	7.943	7.346
w/o writer	7.154	7.983	7.033	6.384
w/o refinement	7.885	8.498	7.704	7.362
w/o agentic	8.153	8.752	8.016	7.622
w/o code	8.217	8.833	8.110	7.655
DeepSurvey	**8.341**	**8.865**	**8.183**	**7.897**

Table 3: Ablation study of content quality when removing each core module (denoted as "w/o").

Method	Recall	Precision	Valid Ratio
w/o collector	0.447	0.406	0.970
w/o keynote	0.600	0.529	0.989
w/o analyzer	0.626	**0.574**	0.998
w/o writer	0.593	0.509	0.610
w/o refinement	0.613	0.537	0.994
w/o agentic	0.591	0.511	**1.0**
w/o code	0.617	0.567	0.995
DeepSurvey	**0.630**	0.569	0.996

Table 4: Ablation study of citation quality when removing each core module (denoted as "w/o").

analysis. In addition, we add Valid Citation Ratio to measure citation authenticity and formatting correctness: citations that cannot be parsed by paper APIs are treated as false references[3].

Tables 3 and 4 show that DeepSurvey's gains come from complementary components rather than any single module. The **collector** is the most critical: its removal causes the largest degradation (total: 8.341 to 6.031; recall/precision: 0.630/0.569 to 0.447/0.406). The **writer** is essential for reliability: its removal collapses valid citation ratio to 0.610. The **keynote** and **analyzer** modules mainly contribute to depth. Removing the analyzer slightly improves precision (0.569 to 0.574) but substantially reduces depth, revealing a mild depth–reliability trade-off that the full system balances via coordinated design. The **refinement** stage brings consistent gains to both content and citation metrics. The simple skill-loop variant (**w/o agentic**) consistently underperforms the central planning agent on both metrics, indicating that globally coordinating review and revision with draft state, evidence, and revision history is more effective than locally iterating without shared planning context. The scores of detailed dimensions in the ablation experiment are in Appendix E.4. Other ablation experiments are supplemented in Appendix E.

Code Agent Subsystem's Gain.

Since code repositories are sparse in non-CS fields, we further compare DeepSurvey with DeepSurvey-noCode in the CS domain on DSEval, together with two advanced dimensions tailored to CS surveys: **Implementation Fidelity** (reflecting implementation details) and **Practical Reproducibility** (guiding the reproduction and improvement of baselines); detailed criteria are in Appendix D.3. As shown in Figure 6, the code agent's gain concentrates in the CS domain ($+0.18$/$+0.41$ in Total Score/Content Depth versus $+0.12$/$+0.24$ across all domains in Table 3), with pronounced improvements on the advanced dimensions (IF and PR improve by 12.4% and 18.5%).

3.6 META-EVALUATION

Meta-evaluation is performed to examine the reliability of the evaluation protocol. First, for reliability of LLM-judge, repeated evaluations exhibit low coefficients of variation (CV) (Pearson, 1897;

[3]Due to the timeliness of the content in the paper's API and issues with API response, the value may contain errors (within 1%).

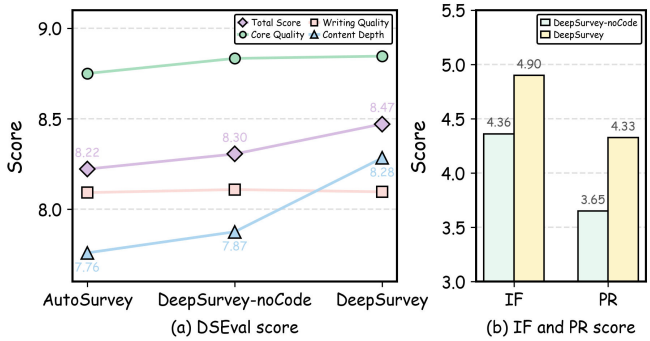


Figure 6: Comparison of AutoSurvey, DeepSurvey-noCode, and DeepSurvey in the CS domain: (a) DSEval scores; (b) Implementation Fidelity (IF) and Practical Reproducibility (PR) scores.

Reed et al., 2002; Shoukri et al., 2006), with *MiMo-no-exp* reaching 0.244%, well below the 5% threshold (Table 6; Appendix B.2). These results confirm that the LLM judge is a reliable component of the evaluation protocol. Second, for LLM–human judge agreement, the LLM judge achieves a mean Cohen's κ (Cohen, 1960) of 0.646, close to the human–human agreement (0.696). At the group level, replacing one human with the LLM yields a Fleiss' κ (Fleiss, 1971) of 0.640, comparable to the all-human panel (0.658) and far above the random-replacement baseline (0.246) (Figure 7; details in Appendix B.3).

4 Related Work

4.1 Traditional Literature Summarization

Traditional literature summarization builds on content selection and citation-based relation modeling. Early extractive methods progressed from frequency-based heuristics (Luhn, 1958) and trainable models (Kupiec et al., 1995) to graph-based centrality such as LexRank (Erkan & Radev, 2011). Citation indexing (Garfield, 1955) and co-citation analysis (Small, 1973) revealed that citation patterns encode field structure, and citation-aware summarization (Cohan & Goharian, 2015; Wang et al., 2018) combined content understanding with inter-paper relation modeling.

4.2 LLM-Based Survey Generation

Recent LLM-based methods decompose survey generation into retrieval, planning, and writing stages. AutoSurvey (Wang et al., 2024b) established the paradigm of semantic retrieval with parallel generation, extended by AutoSurvey2 (Wu et al., 2025) and LLM×MapReduce (Zhou et al., 2025). Subsequent work addressed specific dimensions: SurveyForge (Yan et al., 2025) introduced memory-augmented retrieval; SurveyX (Liang et al., 2025) proposed AttributeTree for structured representation; SurveyGen (Bao et al., 2025) incorporated quality-aware feedback; and multi-agent systems (Go et al., 2025; Liu et al., 2025b) explored collaborative generation.

5 Conclusion

We present **DeepSurvey**, an agentic survey generation system built for agent consumers, for whom analytical depth and citation reliability matter more than breadth. It constructs a reusable analysis substrate through graph-based retrieval, full-text keynotes, cross-paper relation modeling, and code analysis, and enforces reliability through hybrid filtering, evidence-constrained analysis/writing, and multi-granularity refinement. Across multiple domains, DeepSurvey achieves the best content quality (8.341/10) and citation quality (0.630 recall, 0.569 precision, 25.3%/35.2% above the strongest baseline), and is preferred over human-written surveys by domain experts. Most notably, a coding agent reading a DeepSurvey survey autonomously outperforms other baselines, suggesting that the depth such surveys convey is what ultimately transfers to research agents.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, assisting with translation, and interpreting results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, assisting in the writing of proofs, implementing methods, or for cleaning and reformatting datasets, and the remaining required-disclosure tasks (generating synthetic data sets, formulating mathematical claims and providing critical ingredients for proving mathematical claims) are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our human expert evaluation (Section 3.4, Appendix D.7) involved only voluntary participants recruited from students and researchers with relevant academic experience; no crowdworkers were used. All annotations were performed in a blinded setting. Participants were informed that the data would be used exclusively for research and reported only in aggregate, and no personally identifying information was collected or will be released. All papers used in retrieval and evaluation are obtained through public academic services (Semantic Scholar, arXiv) in accordance with their terms of service. Human-written surveys are used solely for evaluation and are excluded from generation to prevent data leakage.

DeepSurvey is intended as an assistive tool for literature synthesis, not a means of producing deceptive academic content. Generated surveys are grounded in verifiable, source-attributed evidence, and we explicitly discourage using the system to generate publications or sharing its output without disclosing that it is not human-written. To further support responsible use, we plan to release an annotation tool that flags potentially unsupported claims and citations. We do not foresee other direct risks of harm, as the system neither collects personal data nor makes consequential automated decisions.

REPRODUCIBILITY STATEMENT

The DeepSurvey pipeline is described in Section 2.1–2.5, and the key prompts for clustering, outline planning, and drafting are provided in Appendix H, which together allow the system to be re-implemented in full. All major models in our system are open-source: MiMo-V2.5 serves as the generation backbone (identical across DeepSurvey and all baselines), MiMo-V2.5 as the primary judge for both content and citation evaluation, and MiniLMv2 as the embedding model. Rerunning the pipeline is therefore inexpensive (Appendix D.6) and does not depend on proprietary APIs. The evaluation protocol is specified in detail, including evaluation topics across five domains (Appendix D.1), metric definitions and aggregation weights (Appendix D.2), system and decoding parameters (Appendix D.4, Table 23), and judge-model validation against human experts (Appendix B.2, B.3). All baselines are run with their official implementations under documented settings. Main results are reported as mean and standard deviation over repeated runs, with effect sizes given in Appendix D.5. The complete source code is available in an anonymous repository (`https://anonymous.4open.science/r/DeepSurvey-anonymous/`) and is also submitted with the supplementary material; the full repository, together with the prompts and the evaluation annotation tool, will be publicly released upon publication.

REFERENCES

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. SurveyGen: Quality-aware scientific survey generation with large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2712–2736, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.136. URL `https://aclanthology.org/2025.emnlp-main.136/`.

Yu Chao, Siyu Lin, Xiaorong Wang, Zhu Zhang, Zihan Zhou, Haoyu Wang, Shuo Wang, Jie Zhou, Zhiyuan Liu, and Maosong Sun. Llm×mapreduce-v3: Enabling interactive in-depth survey generation through a mcp-driven hierarchically modular agent system. In *Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://api.semanticscholar.org/CorpusID:282057493.

Wei-Fan Chen, Zhixue Zhao, Akbar Karimi, and Lucie Flek. Explainable hallucination through natural language inference mapping. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1888–1896, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.96. URL https://aclanthology.org/2025.findings-acl.96/.

Arman Cohan and Nazli Goharian. Scientific article summarization using citation-context and article's discourse structure. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 390–400, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1045. URL https://aclanthology.org/D15-1045/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960. URL https://api.semanticscholar.org/CorpusID:15926286.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, JingChang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jun-Mei Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, Rui-Qi Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shao-Ping Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, T. Sun, Wangding Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen xuan Yu, Wentao Zhang, X. Q. Li, Xiangyu Jin, Xianzu Wang, Xiaoling Bi, Xiaodong Liu, Xiaohan Wang, Xi-Cheng Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yao Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yi-Bing Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxiang Ma, Yuting Yan, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-Long Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437, 2024. URL https://api.semanticscholar.org/CorpusID:275118643.

DeepSeek-AI, Anyi Xu, Bang Lin, Bing Xue, Bing-Li Wang, Bin Xu, Bo Wu, Bowei Zhang, Chao Lin, Chengyao Dong, Chen Ling, Cheng-Da Lu, Chen Zhao, Chengqi Deng, Cheng Hou, Chen Xu, Chen-Ze Shao, Chong Ruan, Conner Sun, Da-Mai Dai, Daya Guo, De-Bin Yang, Deli Chen, Dong-Hui Li, Dong-Li Ji, Erhang Li, Fang-Chen Wei, Fangyun Lin, Fang Yuan, Fei Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guo Cao, Guo-Hui Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Hao Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Hao Yuan, Haowei Zhang, Hao-Wen Luo, Hao Chen, Haozhe Ji, Heng Zhang, Honghui Ding, Hongxuan

Tang, Huanqi Cao, Hua-Zuo Gao, Hui Qu, Hui Zeng, J. Yang, J.-Q. Zhu, Jianming Luo, Jia Song, Jia Yu, Jialiang Huang, Jia-Lu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jia-Shi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jing-Chang Chen, Jing Zhou, Jing Xiang, Jingyang Yuan, Jing Cheng, Jing-Zi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Junliang Ran, Jun Jiang, Jun-Jie Qiu, Jun-Long Li, Junming Zheng, Jun-Mei Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lin Luo, Lin Ma, Linyan Zhu, Li-Tong Wang, Li-Yu Cai, Liyue Zhang, Long-Hao Chen, Mingxi Di, My Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Min Han, Ning Wang, Pan Huang, Pan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qing-Yang Li, Qinyu Chen, Qiushi Du, Qi Jiang, Rui Tian, Rui qin Xu, Rui-Jie Lu, Ruiling Xu, Ruiqi Ge, Rui-Song Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruo-Han Shen, Ruoyu Zhang, Ruyi Chen, Sh. L. Liu, Shanghao Lu, Shangmian Sun, Shang-Yan Zhou, Shan-Shan Chen, Shaofei Cai, Shao Nie, Shao-Ping Wu, Shao-Yuan Chen, Shengding Hu, Sheng Liu, Shiqiang Hu, Shi-Rong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian-Hong Pei, Tian Ye, Tian-Le Lin, Tian Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tu-Tu Wang, W. Y. Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wen-Feng Liang, Wenjie Pang, Wenjing Luo, Wenjin Yao, Wenjun Gao, Wenkai Yang, Wen-lve Huang, Wenqing Hou, Wentao Zhang, Wen-Ting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xing-Chen Liu, Xing-Kai Yu, Xing-You Li, Xin-Yu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xue-Yin Chen, Xuheng Lin, Xu Fu, YC Yan, YQ Wang, Yw Ma, Yanfen Luo, Yang Zhang, Yan-Hong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Xu, Yao Zhao, Yao-Feng Sun, Yao-Hui Wang, Yi Qian, Yingxia Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yi-Jia Wu, Yi-Yu Xiong, Yi Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yi-Xuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yong-Lu Yang, Yong-Qiang Guo, Yong-Tong Wu, Yu Wu, Yukun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yue Xu, Yuhan Wu, Yu Meng, Yu Zou, Yukun Zha, Yunfan Xiong, Yu-Peng Chen, Yuping Lin, Yu Cao, Yuqian Wang, Yushun Zhang, Yu-Ting Yan, Yu-Tong Lin, Yu-Xian Gu, Yu-Wei Luo, Yu-Mei You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, Zhanghua Wu, Ze Wang, Ze Zhao, Zehui Ren, Ze-Kai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhen-Da Xie, Zhengyan Zhang, Zhe Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhi-Xiang Huang, Zhi-Xuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zi jia Zhu, Zi-Long Li, Zipeng Zhang, Zi-Wei Xie, Ziyi Gao, Zi-Yi Wan, Zi-Zheng Pan, and Zong xiang Yao. Deepseek-v4: Towards highly efficient million-token context intelligence. 2026. URL https://api.semanticscholar.org/CorpusID:289623513.

Yufeng Du, Minyang Tian, S. Ronanki, Subendhu Rongali, S. Bodapati, A.G. Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. Context length alone hurts llm performance despite perfect retrieval. *ArXiv*, abs/2510.05381, 2025. URL https://api.semanticscholar.org/CorpusID:281826429.

Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR*, abs/1109.2128, 2011. URL http://arxiv.org/abs/1109.2128.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971. URL https://api.semanticscholar.org/CorpusID:143544759.

Eugene Garfield. Citation indexes for science. *Science*, 122(3159):108–111, 1955. doi: 10.1126/science.122.3159.108. URL https://www.science.org/doi/abs/10.1126/science.122.3159.108.

Gregory Hok Tjoan Go, Khang Ly, Anders Sogaard, Amin Reza Tabatabaei, Maarten de Rijke, and Xinyi Chen. Lira: A multi-agent framework for reliable and readable literature review generation. In *AAAI Conference on Artificial Intelligence*, 2025. URL https://api.semanticscholar.org/CorpusID:281886200.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, and Jun Zhu. Pinnacle: A comprehensive benchmark of physics-informed neural networks for solving pdes, 2023. URL https://arxiv.org/abs/2306.08827.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *ArXiv*, abs/2405.01535, 2024. URL https://api.semanticscholar.org/CorpusID:269502688.

Julian Kupiec, Jan O. Pedersen, and Francine R. Chen. A trainable document summarizer. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995. URL https://api.semanticscholar.org/CorpusID:5775833.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Simin Niu, Shichao Song, Zehao Lin, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. Surveyx: Academic survey automation via large language models. *ArXiv*, abs/2502.14776, 2025. URL https://api.semanticscholar.org/CorpusID:276482768.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023a. URL https://api.semanticscholar.org/CorpusID:259360665.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Ann John, Yassine Benajiba, and Dan Roth. Towards long context hallucination detection. In *North American Chapter of the Association for Computational Linguistics*, 2025a. URL https://api.semanticscholar.org/CorpusID:278165123.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yixin Liu, Yonghui Wu, Denghui Zhang, and Lichao Sun. Agentic autosurvey: Let llms survey llms. *ArXiv*, abs/2509.18661, 2025b. URL https://api.semanticscholar.org/CorpusID:281495800.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *ArXiv*, abs/2408.06292, 2024. URL https://api.semanticscholar.org/CorpusID:271854887.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2:159–165, 1958. URL https://api.semanticscholar.org/CorpusID:15475171.

Kuang Luo, Jing-Shang Zhao, Ying-Ping Wang, Jiayao Li, Junjie Wen, Jiong Liang, Henry Soekmadji, and Shao-Lin Liao. Physics-informed neural networks for pde problems: a comprehensive review. *Artificial Intelligence Review*, 58, 2025. URL https://api.semanticscholar.org/CorpusID:280401971.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. Lost in inference: Rediscovering the role of natural language inference for large language models. *ArXiv*, abs/2411.14103, 2024. URL https://api.semanticscholar.org/CorpusID:274165430.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896, 2023. URL https://api.semanticscholar.org/CorpusID:257557820.

Minh-Anh Nguye, Minh Duc Nguyen, Ha Lan N.T., Kieu Hai Dang, Nguyen Thanh Dong, and Dung D. Le. Surveyg: A multi-agent llm framework with hierarchical citation graph for automated survey generation. *ArXiv*, abs/2510.07733, 2025. URL https://api.semanticscholar.org/CorpusID:281951129.

Karl Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498, 1897.

George F. Reed, Freyja Lynn, and Bruce D. Meade. Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Vaccine Immunology*, 9:1235 – 1239, 2002. URL https://api.semanticscholar.org/CorpusID:231319.

Xiaofeng Shi, Qian Kou, Yuduo Li, Ning Tang, Jinxin Xie, Longbin Yu, Songjing Wang, and Hua Zhou. Scisage: A multi-agent framework for high-quality scientific survey generation. *ArXiv*, abs/2506.12689, 2025. URL https://api.semanticscholar.org/CorpusID:279402998.

Mohamed M Shoukri, Nasser Elkum, and Stephen D. Walter. Interval estimation and optimal design for the within-subject coefficient of variation for continuous and binary variables. *BMC Medical Research Methodology*, 6:24 – 24, 2006. URL https://api.semanticscholar.org/CorpusID:264815049.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris,

Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Y. Guan, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomek Korbak, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, July 1973. doi: 10.1002/asi.4630240406. URL https://ideas.repec.org/a/bla/jamest/v24y1973i4p265-269.html.

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Xuanyi Chen, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. Surge: A benchmark and evaluation framework for scientific survey generation. *arXiv preprint arXiv:2508.15658*, 2025.

Alex D Wade. The semantic scholar academic graph (s2ag). *Companion Proceedings of the Web Conference 2022*, 2022. URL https://api.semanticscholar.org/CorpusID:251597885.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction, 2024a. URL https://arxiv.org/abs/2409.18839.

Haoyu Wang, Yujia Fu, Zhu Zhang, Shuo Wang, Zirui Ren, Xiaorong Wang, Zhili Li, Chaoqun He, Bo An, Zhiyuan Liu, and Maosong Sun. Llm×mapreduce-v2: Entropy-driven convolutional test-time scaling for generating long-form articles from extremely long resources, 2025. URL https://arxiv.org/abs/2504.05732.

Jie Wang, Chengzhi Zhang, Mengying Zhang, and Sanhong Deng. Citationas: A tool of automatic survey generation based on citation content. *Journal of Data and Information Science*, 3:20–37, 06 2018. doi: 10.2478/jdis-2018-0007.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *ArXiv*, abs/2012.15828, 2020. URL https://api.semanticscholar.org/CorpusID:229923069.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. Autosurvey: Large language models can automatically write surveys. *ArXiv*, abs/2406.10252, 2024b. URL https://api.semanticscholar.org/CorpusID:270560509.

Siyi Wu, Chia Xin Liang, Ziqian Bi, Leyi Zhao, Tianyang Wang, Jun-Jie Song, Yichao Zhang, Keyu Chen, and Xinyuan Song. Autosurvey2: Empowering researchers with next level automated literature surveys. *ArXiv*, abs/2510.26012, 2025. URL https://api.semanticscholar.org/CorpusID:282592838.

Xi Xiao, Bing Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hong-Mei Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun-Miao Shi, Jun Jacob Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qian Chen, Shao yang Liu, Shi liang Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tian-Yu Zhou, Wei Su, Weikun Wang, Wenhan Ma, Xia Deng, Bo Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fang Liu, Feiyu Yang, Feng Shi, Guoan Wang, Hao Tian, Hao Wu, Hengxu Qu, Hong Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yi-Hao Song, Yi-Tong Yan, Yihao Zhao, Ying Lai, Yizhao Gao, Yu Cheng, Yu Tian, Yudong Wang, Zheng-Yu Tang, Zheng-Yu Tang, Zheng Wen, Zhichao Song, Zhixian Zheng, Zi-Ang Jiang, Jiantao Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shao-Qiang Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shu-Yue Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wen-Jie Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yuanyu Tu, Yu-Jie Dong, Yuchen Liu, Yue Ma, Yue Yu, Yu-Cui Xiang, Zhaojun Huang, Zhenrui Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report. *ArXiv*, abs/2601.02780, 2026. URL https://api.semanticscholar.org/CorpusID:284513060.

Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. *Proceedings of the 26th International Conference on World Wide Web*, 2017. URL https://api.semanticscholar.org/CorpusID:1644335.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *ArXiv*, abs/2504.08066, 2025. URL https://api.semanticscholar.org/CorpusID:277741107.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:276813240.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023. URL https://api.semanticscholar.org/CorpusID:259129398.

Zihan Zhou, Chong Li, Xinyi Chen, Shuo Wang, Yu Chao, Zhili Li, Haoyu Wang, Qi Shi, Zhixing Tan, Xu Han, Xiaodong Shi, Zhiyuan Liu, and Maosong Sun. LLM×MapReduce: Simplified long-sequence processing using large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27664–27678, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1341. URL https://aclanthology.org/2025.acl-long.1341/.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *ArXiv*, abs/2310.17631, 2023. URL https://api.semanticscholar.org/CorpusID:264490588.

A DeepSurvey System Supplement

A.1 Code Analysis Subsystem

Beyond paper text, DeepSurvey analyzes open-source code repositories linked in collected papers to supplement implementation details not fully described in the text.

Pseudocode generation. The code analysis subsystem downloads repositories and produces repository-level pseudocode through a multi-agent loop centered on a planning agent. At each step, the planner selects one of five operations: GET_SOURCE_CODE reads a specific file from the repository; CREATE generates initial pseudocode from retrieved context; REVISE refines the pseudocode based on reviewer feedback; REVIEW invokes a reviewer that scores the current pseudocode on conciseness, logical structure, and implementation specificity (each 0–10) with concrete improvement suggestions; and FINISH terminates the loop. The planner is required to read 3–5 key source files (entry points, core algorithm modules, key utilities) before creating pseudocode, and to call REVISE at least once every 3 rounds. For the integrity of the plan, the planner gives several steps (a sub-plan) at a time rather than one step. A memory mechanism tracks operation history and review feedback across steps, with automatic compression when token count exceeds a threshold to prevent context overflow.

Report generation. The system produces two types of reports. **Code reports** are generated by processing pseudocode in batches (batch size 5): each batch is analyzed along dimensions including problem modeling, core algorithm classification, optimization strategies, and topic-specific dimensions (e.g., agent architectures for LLM systems); batch reports are then integrated into a final report, with multi-round fallback when combined content exceeds the context window. The prompt enforces traceability—each claim must reference specific pseudocode components. **Environment reports** extract configuration files (requirements.txt, README) from repositories, analyzing framework choices, dependency versions, and deployment patterns. Both reports use semi-open prompts that mandate coverage of core dimensions (algorithm classification, data modeling, engineering optimizations) while allowing the model to identify topic-specific insights.

Exception Handling. There are always extreme cases in a large number of code repositories, and DeepSurvey is capable of handling exceptions. For **large repositories**, the agent progressively explores the repo structure and reads core files step by step, which handles large repos as long as their size stays within the agent's analysis capacity and we additionally apply filtering (e.g., excluding cache directories, outputs, and build artifacts). For extreme cases, we truncate the repo structure. For **multi-repo projects**, all open-source repos are analyzed individually and then jointly processed in the second stage. Users can configure the system to analyze only the first **k** repos; in our experiment, we read all the provided repos. For **missing/obsolete repos**, repos with missing files or inaccessible URLs are filtered out with tools before analysis.

A.2 Analysis Knowledge Base

The analysis knowledge base serves as the central artifact connecting DeepSurvey's understanding and generation stages. Rather than treating survey generation as a one-shot text production task,

DeepSurvey first constructs a structured, reusable knowledge repository from the collected literature, then grounds all subsequent writing and refinement in this repository. This design separates *analysis* from *generation*, enabling deeper reasoning without overwhelming the writing model.

The knowledge base is organized into three hierarchical levels, each building on the previous:

- **Paper level**: Full-text *keynotes* recording core mechanisms, experimental settings, key assumptions, applicable ranges, and limitations for each paper. Additionally, for papers with linked code repositories, *code reports* and *environment reports* are generated to capture implementation-level details such as module decomposition, execution flow, engineering dependencies, and runtime configurations—information that is typically absent from paper text.
- **Cluster level**: Papers are organized into thematic clusters, each annotated with three complementary analysis views: *relation graphs* that model typed citation relationships (foundation, extension, substitution) with abstracted descriptions; *comparison tables* that conduct structured cross-paper comparison along key technical dimensions; and *guided Q&A analyses* that integrate evidence to answer high-value research questions within and across clusters.
- **Survey level**: Hierarchical outlines, section drafts, and revision plans that directly serve the generation process.

These representations are linked through explicit mappings: paper keynotes reference their cluster assignments, cluster analyses cite specific papers, and section drafts trace their evidence to both paper keynotes and cluster analyses. This traceability supports the verification mechanisms throughout the pipeline and enables the iterative refinement subsystem to identify and correct inconsistencies.

The knowledge base is designed with two complementary objectives. First, it serves as the evidential foundation for survey generation: the outline-driven writing module assigns papers to sections based on cluster analysis results, and the refinement module retrieves keynotes and code reports to verify and deepen individual paragraphs. By grounding generation in structured intermediate representations rather than raw retrieval results, the system maintains content depth while constraining hallucination. Second, the knowledge base constitutes an independently valuable research artifact. The paper-level keynotes, cluster-level relation graphs, and comparison tables provide fine-grained, traceable insights into a research landscape—supporting tasks such as literature reviews, method comparisons, and research direction analysis beyond the immediate scope of survey writing. This dual utility reflects our design philosophy: the analysis knowledge base is not merely a byproduct of survey generation, but a reusable substrate for scientific inquiry.

A.3 Engineering Details

To support large-scale processing, DeepSurvey incorporates several engineering optimizations.

Multi-domain support. DeepSurvey uses fallback mechanisms to improve robustness across domains. When Semantic Scholar returns insufficient relevant papers or fails during retrieval, the system invokes the arXiv API to supplement the evidence pool. When full texts are unavailable, which is a common phenomenon in certain fields, DeepSurvey falls back to abstracts or TLDRs as lightweight substitutes for full-text keynotes, preserving basic paper coverage while allowing downstream analysis to proceed.

Unified paper identifier. DeepSurvey retrieves papers from multiple sources (Semantic Scholar, arXiv), each with its own identifier system. To ensure consistent caching and deduplication across the pipeline, we construct a unified `paper_id` for each paper: if an arXiv ID is available, it serves as the canonical identifier; otherwise, the Semantic Scholar ID is used as a fallback. All downstream caches (keynotes, relation graphs, analysis results) are hash-keyed by this unified `paper_id`, enabling cross-session reuse and preventing redundant processing of the same paper retrieved from different sources.

Robust paper downloading. Paper downloads support streaming writes and resumable transfer through HTTP range requests. After downloading, PDF files are validated with `pypdfium2`; corrupted or incomplete files are removed and retried to ensure reliable downstream parsing.

Multi-level caching. A three-tier caching system minimizes redundant computation and API costs: (1) *API-level caching* stores raw responses from external services (Semantic Scholar, LLMs, embedding models) keyed by request parameters, with configurable expiration; (2) *Task-level caching* persists results of major pipeline stages (literature collection, keynotes, clustering, analysis) as JSON files for session-level reuse; (3) *Runtime caching* maintains in-memory results within pipeline stages to eliminate intra-stage redundancy. Task-level caching is implemented via `diskcache`, enabling cross-session reuse of expensive computations such as full-text keynotes and relation graphs.

Parallel processing. Independent operations execute concurrently via `ThreadPoolExecutor`: keynote generation for different papers, comparison table construction across clusters, and batch LLM API calls all proceed in parallel. Thread pool sizes are configured per task type to balance throughput against API rate limits. Each request includes timeout control to prevent individual failures from blocking the pipeline. A checkpoint mechanism enables resuming interrupted pipelines from the last completed stage rather than restarting from scratch.

Error recovery. LLM API calls are wrapped with automatic retry via the `tenacity` library, using exponential backoff (1–300 s) for up to 10 attempts on recoverable HTTP errors (408, 429, 500, 502, 503, 504). At the pipeline level, a structured error memory accumulates recent failure messages (capped at 10 entries with deduplication) and injects them into subsequent prompts, enabling the model to adapt its output based on prior failure patterns. This mechanism is particularly effective in multi-round tasks such as outline generation and citation assignment.

Document parsing. Full-text analysis relies on MinerU (Wang et al., 2024a) to convert PDFs into structured Markdown while preserving layout, reading order, and mathematical notation. Each paper is parsed into an independent output directory containing the structured text and intermediate layout annotations, which serve as input to the full-text understanding module.

Context window management. Given LLM context limits, DeepSurvey employs adaptive input construction: token counts are estimated via `tiktoken` (falling back to the `cl100k_base` encoder), and inputs exceeding the model's window are progressively compressed—first by trimming retrieval-augmented context (halved iteratively), then by truncating paper content while preserving core keynotes and outline structure. In the citation assignment stage, this progressive reduction ensures that essential paper–outline mappings are retained even under tight context constraints.

Device adaptation. The embedding model uses lazy loading with automatic GPU-to-CPU fallback on out-of-memory errors. Batch encoding sizes are configurable to balance throughput against memory footprint during large-scale similarity computation.

Agent memory management. For multi-step agentic tasks, DeepSurvey stores operation history, intermediate outputs, review feedback, and revision states in structured memory. When the token budget is exceeded, older long-context entries are compressed or truncated while recent actions and a compact state summary are preserved for subsequent planning.

A.4 Reliability Check

Although DeepSurvey's citation reliability significantly surpassed baselines, citation issues still occur at a relatively low frequency. Since our motivation is to develop a research tool for both researchers and agents, we expect DeepSurvey's outcome readers to themselves be DeepSurvey users, who can judge survey quality by combining DSEval results with their own knowledge.

We also plan to release an annotation tool as part of our open-source repository that examines citation reliability for each claim based on DSEval's per-claim citation evaluation method and flags suspicious claims and citations, helping both human and AI users use DeepSurvey more reliably.

Table 5: Scores under AutoSurvey's original evaluation protocol. Both AutoSurvey and DeepSurvey saturate at the maximum score (5), indicating insufficient discriminative power.

Method	Coverage	Structure	Relevance
AutoSurvey	5	5	5
DeepSurvey	5	5	5

We do not currently encourage any publication via DeepSurvey, or sharing its output with others without disclosing that it is not human-written.

B Evaluation Protocol Validation Supplement

B.1 AutoSurvey Evaluation Experiment

AutoSurvey (Wang et al., 2024b) established the dominant evaluation paradigm for automated survey generation, assessing content quality across three dimensions—Coverage, Structure, and Relevance—on a 1–5 scale using LLM-as-judge. Its evaluation framework has been widely adopted by subsequent work (Wu et al., 2025; Wang et al., 2025; Liang et al., 2025; Yan et al., 2025; Go et al., 2025; Shi et al., 2025), and its original meta-evaluation demonstrated moderate positive correlation with human judgments (Spearman's ρ up to 0.5429), confirming its validity at the time.

To assess whether this benchmark remains discriminative under current model capabilities, we replicate AutoSurvey's evaluation protocol exactly—using its original dimensions, prompts, and scoring criteria—to evaluate AutoSurvey and DeepSurvey. As shown in Table 5, both methods achieve the maximum score (5) across all dimensions, rendering the benchmark unable to differentiate between approaches.

This saturation arises because AutoSurvey's 5-point scale and coarse-grained dimensions were designed for an earlier generation of models. As LLM capabilities have advanced, state-of-the-art systems—regardless of their architectural differences—now consistently meet the ceiling criteria for basic coverage, structural coherence, and topic relevance.

B.2 LLM-Judge Stability Experiment

LLM-as-judge has emerged as a scalable alternative to human evaluation for open-ended text quality assessment. Prior work has shown that strong judge models (e.g., GPT-4) achieve over 80% agreement with human preferences (Zheng et al., 2023), and frameworks such as G-Eval (Liu et al., 2023b) demonstrate reliable human correlation on summarization tasks. However, LLM judges are not inherently stable—JudgeLM (Zhu et al., 2023) and subsequent studies on position bias (Kim et al., 2024) show that scoring reliability is sensitive to prompt format, answer ordering, and task configuration. It is therefore necessary to explicitly validate judge stability before relying on LLM scores for cross-method comparison.

We evaluate four judge configurations along two axes: **model** (GPT series vs. MiMo series) and **prompt style** (with explanation, `-exp`, vs. without explanation, `-no-exp`). For each configuration, we score the same set of generated surveys multiple times and compute per-sample standard deviation, coefficient of variation (CV%), maximum absolute deviation, and pairwise maximum difference. Aggregated results are reported in Table 6; fine-grained breakdowns by metric and method are provided in Table 20. Note that we do not perform a separate judge-stability study for the NLI-based citation evaluation because prior work has shown that NLI remains informative for LLM evaluation and is widely used for fact verification and hallucination detection (Manakul et al., 2023; Madaan et al., 2024; Chen et al., 2025; Wang et al., 2024b).

Table 6: Aggregated stability statistics across four judge configurations. "-exp" indicates that the LLM must give a score with an explanation, while "-no-exp" indicates the reverse.

Configuration	std	CV%	max_abs_dev	range
GPT-exp	0.041	0.49	0.044	0.081
GPT-no-exp	0.023	0.279	0.024	0.046
MiMo-exp	0.077	1.03	0.078	0.152
MiMo-no-exp	**0.020**	**0.244**	**0.022**	**0.035**

Results. The `mimo-no-exp` configuration achieves the best stability across all four metrics (std = 0.020, CV = 0.244%, max_abs_dev = 0.022, pairwise_max_diff = 0.035), representing absolute fluctuations below 5%—well within acceptable bounds for cross-method ranking. Compared to `gpt-no-exp`, `mimo-no-exp` reduces standard deviation by 13.0% and CV by 12.5%. The improvement over `mimo-exp` is even more pronounced: std, range, CV, and max_abs_dev decrease by 74.0%, 77.0%, 76.3%, and 71.8% respectively (range and pairwise_max_diff refer to the same statistic), indicating that requiring the judge to produce explanations introduces substantial additional variance that undermines scoring consistency.

Choice of MiMo-series model as judge model. We select `mimo-no-exp` (MiMo models without explanation prompts) as the primary judge for our evaluation setup for two reasons. First, it offers the highest scoring stability among all candidates, ensuring that observed differences between methods reflect genuine quality distinctions rather than judge noise. Second, MiMo is an open-source series model with relatively low inference cost, enabling reproducible and scalable evaluation without dependence on proprietary APIs. The detailed per-metric stability breakdowns in Table 20 further confirm that `mimo-no-exp` maintains consistently low variance across different methods (AutoSurvey, DeepSurvey) and evaluation dimensions (Core Quality, Writing Quality, Content Depth), validating its suitability as a unified evaluation standard.

B.3 Consistency Between LLM Judge And Human Experts Experiment

We report agreement at both the pairwise and group levels. Pairwise agreement is measured with Cohen's κ, and panel-level consistency is measured with Fleiss' κ. Since our labels are nominal (A, B, and Tie), every item has a fixed number of annotations, and there are no missing judgments, nominal Krippendorff's α and Fleiss' κ are numerically very close in our setting, with the difference arising only from their finite-sample treatment of chance agreement. We therefore report Fleiss' κ as the group-level metric to avoid redundant notation.

Cohen's κ. For two annotators, Cohen's κ is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

where p_o is the observed agreement and p_e is the chance agreement implied by the annotators' marginal label distributions. We compute Cohen's κ between the LLM judge and each human expert, and between pairs of human experts, then average over all valid pairs.

Fleiss' κ. For multi-rater agreement, let n_{ic} be the number of annotators who assign category c to item i, and let n be the number of annotators per item. Fleiss' κ is computed as

$$P_i = \frac{1}{n(n-1)} \sum_c n_{ic}(n_{ic} - 1), \tag{2}$$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i, \tag{3}$$

$$p_c = \frac{1}{Nn} \sum_{i=1}^{N} n_{ic}, \tag{4}$$

$$P_e = \sum_c p_c^2, \tag{5}$$

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e}. \tag{6}$$

Here, P_i measures agreement on item i, $\bar{P}$ is the average observed agreement, and P_e is the agreement expected by chance from the marginal label frequencies.

Evaluation settings. We evaluate several group configurations to separate human consistency, LLM contribution, and random baselines:

- **Human full.** All human annotators are included. This setting measures the intrinsic consistency of the human panel and serves as the main human reference.
- **Human leave-one-out.** We remove one human annotator at a time and compute group agreement on the remaining human panel. This setting estimates the natural variation of the human panel under a one-rater perturbation.
- **Human leave-one-out + actual LLM judge.** We replace the removed human with one LLM judge and recompute group agreement. This is the main test of whether the LLM can substitute for one human without degrading panel consistency.
- **Human leave-one-out + random human judge.** We replace the removed human with another human judge sampled from the human pool. This setting provides an empirical human-replacement baseline and approximates the agreement level expected when the missing annotator is replaced by a genuine human.
- **Human leave-one-out + random evaluator.** We replace the removed human with a random evaluator that assigns A, B, or Tie with equal probability $1/3$. This is a null baseline that estimates the agreement expected from an uninformative judge.
- **LLM full.** All LLM judges are included. This setting measures the internal consistency of the LLM panel and serves as a reference for LLM-to-LLM agreement.

Results. The results are shown in Figure 7. The pairwise results show that the LLM judge is close to human experts in fine-grained agreement. Its mean Cohen's κ of 0.6463 is only 0.0495 below the human–human agreement of 0.6958, while remaining far above the human–random baseline (-0.0016). This gap indicates that the LLM judge captures human preference patterns rather than behaving like a random scorer.

The group-level results lead to the same conclusion. Human-1 + LLM achieves a Fleiss' κ of 0.6398, which is close to the all-human panel (0.6578). In contrast, human-1 + random drops sharply to 0.2462, showing that the gain from the LLM is not explained by a generic additional label source. The difference between human-1 + LLM and human-1 + random is large enough to suggest that the LLM contributes structured, human-aligned signal to the panel.

Overall, these numbers support two claims. First, the LLM judge is well aligned with human experts at the pairwise level. Second, when inserted into a multi-rater panel, it preserves overall consensus at a level comparable to human annotators. This makes the LLM judge a stable and credible component of the evaluation protocol.

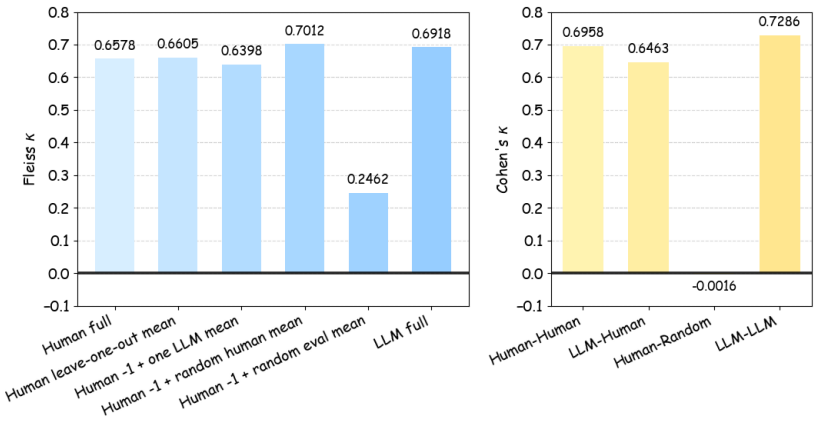


Figure 7: Agreement between human and LLM-judge. The left panel shows Fleiss' κ for compositions including full humans, leave-one-human-out, replacing one human with an LLM, and random-replacement baselines. The right panel shows pairwise Cohen's κ between human, LLM, and random evaluator groups.

C Downstream Validation with a Coding Research Agent

C.1 Setting.

To complement rubric-based scoring with a direct behavioral measure, we examine whether a survey improves the research capability of a downstream agent on PINNacle (Hao et al., 2023), a benchmark for physics-informed neural networks (PINNs). We strip all published PINN optimization methods from the benchmark, leaving only the vanilla baseline (a tanh fully-connected network trained with Adam on the standard losses), so that methodological knowledge must originate from the agent's prior knowledge or the provided survey. We compare four conditions: *none* (no literature input), *human* (a human-written survey), *SurveyForge* (a survey generated by the strongest baseline system), and *DeepSurvey* (a survey generated by our system). For the human condition we select one of the most highly cited human-written surveys on PINNs published within the last two years (Luo et al., 2025), ensuring that this condition provides sufficiently recent methodological knowledge as well as authoritative and reliable expert analysis.

C.2 Case selection.

PINNacle spans regimes on which, as its own evaluation shows, no single published optimization methodology succeeds across all cases. Designing and validating a methodology effective on every case within one fixed wall-clock session (480min) is therefore infeasible by construction: forced to cover the full suite, the agent could not complete even a single design–implement–validate loop, and all conditions would collapse to a floor effect. We thus fix a four-case subset spanning three canonical PDE families: nonlinear advection–diffusion (Burgers-1D-C), parabolic diffusion (Heat-2D-CG, Heat-2D-MS), and hyperbolic propagation (Wave-1D-C), forming a graded difficulty profile in terms of the vanilla baseline: nearly solved on Burgers-1D-C (relative L2 error $\sim 10^{-2}$), moderately difficult on the Heat cases, and severely failed on Wave-1D-C ($\sim 6 \times 10^{-1}$). This gradient allows us to examine whether the effect of survey quality scales with how much methodological knowledge a case demands. The case set was fixed before any agent runs, and identical cases are used across all conditions.

C.3 Session protocol.

In each run, a coding agent (codex[4]) backed by DeepSeek-V4-Flash (DeepSeek-AI et al., 2026) operates in an isolated copy of the workspace and receives exactly one literature input (a DeepSurvey-

[4] https://openai.com/codex/

generated survey, a SurveyForge-generated survey, a human-written survey, or none), together with the locked evaluation protocol and an anonymized leaderboard of published results that is identical across conditions; SurveyForge serves as the strongest baseline, ranking second to DeepSurvey in both total and depth scores. The agent backbone is deliberately chosen from a model family different from that used to generate the surveys (MiMo-V2.5), ruling out self-preference. The leaderboard contains reference error values only, with method identities anonymized so that the agent cannot infer superior approaches from algorithm names; it indicates attainable accuracy. The agent must autonomously design, implement, and validate a PINN training methodology within a fixed wall-clock session, writing only to a dedicated directory with predefined hooks (loss formulation, architecture, loss weighting, optimizer, and training loop); the harness then re-trains the delivered implementation from scratch on the enabled PDE cases at the benchmark's original protocol (20,000 iterations per case), using evaluation seeds disjoint from those exposed through the workspace's development interface to prevent seed overfitting, and reports the mean relative L2 error per case. We deliberately adopt codex rather than a sophisticated auto-research framework: more complex or less constrained systems introduce additional confounding variables, and on a problem of moderate difficulty an overly capable research agent could succeed independently of the provided survey, robbing the experiment of its ability to attribute outcomes to survey quality. Beyond the relative L_2 error metric, we additionally score the novelty of each delivered methodology with GPT-6 Astra[5]: as one of the strongest LLMs, with pretraining knowledge recent enough to cover the current PINN literature, it can judge comparatively objectively whether a methodology departs from established techniques, a quality that the L_2 error metric alone cannot capture. The complete scoring rubric is shown in Table 7.

Table 7: Scoring rubric for the technical novelty of delivered methodologies (judged by GPT-6 Astra).

Score	Standard
1–2	Essentially identical to an existing published method.
3–4	Minor modification or straightforward combination of existing methods; limited new technical content.
5–6	Noticeable methodological novelty; the idea may follow or extend prior work, but adds meaningful new technical content beyond it.
7–8	Clearly novel core mechanism or formulation; no known published work presents essentially the same method.
9–10	Highly original core idea that introduces a substantially new methodological principle or paradigm.

C.4 Fairness controls.

Three controls ensure fairness and elicit genuine research behavior. First, isolation is dual-enforced: the sandbox blocks network access and out-of-workspace reads, all evaluation code is hash-locked and statically checked, and post-session transcript audits screen for protocol violations, including bypassing learning altogether, such as directly solving the PDE mathematically or delegating to external numerical solvers, which lies outside the permitted modification scope of the predefined hooks; since the reported numbers are re-trained from the pristine codebase, session-time tampering cannot affect the results. Second, a uniform research directive appended to every condition's prompt encourages genuinely novel optimization rather than stacking surveyed components or brute-force parameter search, mirroring realistic auto-research. Third, the session budget is wall-clock rather than token-based, as token spending is confounded by survey length; the budget is sized on the order of a full validation pass of the benchmark, which renders exhaustive re-evaluation loops infeasible and compels the agent to rely on insights distilled from the survey rather than repeated trial-and-error.

C.5 Results.

Table 2 reports the best of three runs per condition, each retrained from scratch at the full 20,000-iteration protocol. We report the best run rather than the mean for two reasons: agent sessions

[5] https://openai.com/index/gpt-6-astra/

occasionally exhibit anomalous failures (e.g., environment errors or early abandonment) whose inclusion would distort the average without reflecting the condition's attainable capability, and in research practice the best attempt within a fixed budget is the quantity of interest, mirroring how benchmark results are reported. The selection rule is applied uniformly across all conditions. PINN-base (pilot) is the vanilla baseline trained by our harness itself under the identical protocol before any agent sessions, and PINNacle SOTA takes, for each case, the best value among the published PINNacle methods (the vanilla column excluded); the account below is grounded in the session transcripts of these runs. Only the DeepSurvey condition improves on the published leaderboard reference: its overall error of 1.53e-02 is $11.2\times$ below PINN-base and $2.5\times$ below the per-case-best reference row (3.81e-02), whereas the other three conditions all remain above the reference.

Headroom analysis. The none condition attains the best per-case error only on Burgers-1D-C and Heat-2D-CG, which are exactly the two low-headroom cases: measured as the ratio of the vanilla baseline to the published SOTA, the attainable headroom is merely $1.6\times$ (Burgers-1D-C) and $1.1\times$ (Heat-2D-CG), versus $8.8\times$ (Heat-2D-MS) and $5.0\times$ (Wave-1D-C) on the remaining two. On the two easy cases, the none condition's improvements over the baseline are correspondingly marginal ($1.6\times$ and $1.4\times$) and, per its session report, amount to conservative engineering tweaks. In contrast, DeepSurvey's gains concentrate on the two high-headroom cases, where it improves over the baseline by $12.9\times$ and $51\times$, surpassing the published SOTA on both, and its delivered method (SCRF-PINN) is judged architecturally novel. Improvements that require methodological knowledge therefore track the amount of knowledge a case demands, which is precisely where a survey's value should manifest.

Baseline Results The three non-DeepSurvey conditions converged to conservative or unstable methods, and their failure modes are informative. The SurveyForge-condition agent selected precisely the two classical directions highlighted by its survey, namely a hard-constraint ansatz that satisfies boundary conditions by construction and adaptive per-point residual weighting. The combination clearly outperformed the baseline at the short development horizon ($\sim$1,000 iterations) but degraded in late training under the full protocol, leaving the run at approximate parity with the vanilla baseline (1.63e-01 overall). Transcript analysis attributes this to a validation-horizon gap: the survey presents both techniques at the concept level, without implementation-level construction details (e.g., the boundary extension function of a hard-constraint ansatz) or failure-mode analysis, so the agent spent most of its budget rediscovering these details by trial-and-error and reached the session time limit without full-protocol validation. The human-survey condition marks the upper bound of the same limitation. Its survey's diagnosis of PDE/BC loss imbalance (§2.3.1) gave the agent a correct entry point, and its first action was to measure per-term loss and gradient imbalance; however, the survey's high-level enumeration of remedies led it to stack adaptive gradient reweighting, Fourier features, and residual resampling as published. The adaptive scheme failed systematically in this setting; unable to reconcile the components within budget, the agent retreated to a minimal design: a static boundary-loss weight of 1000 with a 400-step PDE-loss warm-up on three boundary-dominated cases, with all other cases kept on the exact baseline trajectory. This conservatism, together with full-protocol validation during development (unique among these runs), made the result stable (6.48e-02) but capped its novelty and its gains. The no-survey condition lacked even this diagnostic scaffold: with no prior evidence for ranking candidates, it stacked four conventional tricks at once (adaptive loss weighting, causal curriculum, learning-rate scheduling, and L-BFGS finishing), which made component defects hard to attribute; about a quarter of the session went to isolation and rollback, and 62% of the budget was then spent tuning two hard cases. Its final method reverted to the standard pipeline plus two conservative changes (balanced boundary collocation and a bounded end-of-training L-BFGS polish), beating the baseline on all four cases but by only $1.33\times$. Taken together, these runs suggest that neither parametric knowledge alone nor a survey written for human readers provides sufficient implementation-level understanding of the problem, or scenario-specific guidance on selecting and instantiating mechanisms, for the agents to converge within the session budget on a methodology that is both novel and effective.

DeepSurvey Results In contrast, the DeepSurvey survey organizes knowledge mechanism-first. It opens with a unified diagnosis of PINN training failures (flat-minimum traps, spectral bias, fixed-point attraction; §2.1, §5.4) and maps each candidate method to the failure mode it targets: the σ-bandwidth knob that directly modulates spectral bias (§3.1), the hard-constraint output transform

(§3.2), curriculum regularization (§2.6), and the closed-loop adaptive principle (§4), each accompanied by applicability conditions and failure-mode analysis. Every section thus corresponds to a class of actionable interventions rather than a chronological enumeration of prior work. The session transcript shows this organization being translated into design decisions step by step. Within the first two minutes, the agent covered the full mechanism chapters (§1–§5) through seven targeted extractions and delivered a first method version after roughly twenty minutes. When this version failed through a runaway feedback loop in its closed-loop controller, the agent did not retreat to trivial modifications; based on the understanding obtained from the survey, it turned to measuring the problem data itself and built the per-axis spectral fingerprint σ_{req}, and after diagnosing a plateau-trigger defect and measuring that the anchor function alone already attained a relative L2 error of 0.14, it completed the hard-constraint redesign with zero initialization and a distance-ratio anchor. Two keyword-guided re-reads of the survey ("hard constraint", "loss balancing", "flat trap") preceded the reconstruction of the distance-ratio anchor and the drift-timescale learning-rate decay, which the agent's report explicitly attributes to §2.4. Late in the session, before implementing the Helmholtz gate and the wave reparameterization, the agent searched the survey for "helmholtz", "hyperbolic", and "wave equation", found only two passing mentions and no dedicated technique for hyperbolic problems, and only then invented the symmetric d'Alembert ansatz. This behavior, confirming a coverage gap in the survey before filling it autonomously, is the origin of the Wave-1D-C breakthrough.

The delivered SCRF-PINN is organized around operator-measurement-driven automatic decision-making: a per-axis spectral fingerprint measured from the PDE operator (by probing the operator with quadratic test functions) and from the boundary/initial data, which sizes an anisotropic Fourier-feature layer only where the measured bandwidth requires it; a hard-constraint output transform that enforces Dirichlet conditions exactly, automatically disabled when the measured k^2 term identifies a Helmholtz-type operator, for which the harmonic anchor underlying the constraint is misspecified; and an exponential learning-rate schedule timed to the drift timescale. On the four scored cases, the method reaches 1.36e-02 on Burgers-1D-C (1.54$\times$ below the baseline, within 2.3% of the best result reported on the PINNacle leaderboard, 1.33e-02), 1.18e-02 on Heat-2D-MS (12.9$\times$ below the baseline and below the leaderboard best of 1.73e-02), and 9.54e-03 on Wave-1D-C (51$\times$ below the baseline and an order of magnitude below the leaderboard best of 9.79e-02), while Heat-2D-CG remains at baseline level (2.64e-02). The overall error of 1.53e-02 is the best of all conditions, the Wave-1D-C result is the best across all runs, and the GPT-6 Astra novelty judge scores the delivered methodology markedly higher than the methods delivered under the other three conditions.

The survey's role illustrates the core advantage of DeepSurvey: what it provides is not a menu of methods but a deep understanding of the problem itself. The σ knob exists in the survey (§3.1), yet the quadratic-probe spectral-fingerprint measurement, the plateau-triggered closed loop, and the k^2 gating rule are the agent's own constructions with no direct counterpart in the survey text. Reading the survey thus enabled the agent both to reuse appropriate mechanisms selectively and to innovate beyond them. By contrast, the high-level method enumerations available to the human and SurveyForge conditions led to stacking components as published, and the no-survey condition lacked even a diagnostic entry point; after setbacks, all three reverted to trivial variants rather than attempting genuinely novel designs. The value of a DeepSurvey survey therefore lies not only in deeper and more concrete solutions, but in its analysis deepening the agent's understanding of the domain to the point where original mechanisms beyond the survey text become attainable.

C.6 Qualitative Analysis.

Session transcripts and deliverable reports yield two qualitative observations that complement the quantitative results. First, in an early protocol iteration that asked only for the highest achievable score, without the innovation directive or the PINN-methodology restriction, the DeepSurvey-condition agent leveraged its understanding of the PINN task to recognize that several benchmark cases admit direct mathematical solution, solved them outside the learning framework, and achieved exceptionally low errors (roughly two orders of magnitude below the baseline, surpassing the best published results in the leaderboard), while agents under the other conditions remained occupied with tuning the parameters of classical optimization mechanisms. This contrast indicates that DeepSurvey equips the agent with a deeper understanding of the research topic, an implementation-level understanding rather than a high-level recall of existing methods, which is what enabled the agent to recognize that the benchmark objective could be attained without learning at all. We excluded

this run from all quantitative comparisons as out of scope for the research question, which targets learning-based methodologies, and made this restriction explicit in the upgraded protocol.

Second, agents under the SurveyForge, human-written, and no-survey conditions consistently fell back to stacking established methods: even under the identical innovation directive, their reports describe starting from classical techniques drawn from the survey or from pretraining knowledge, combining multiple optimizations, failing to reconcile their hyperparameters, and retreating to trivial variants. We do not attribute this contrast to the DeepSurvey-condition agent following the innovation directive while the other agents fail to follow it, since the directive was uniform across conditions; we instead interpret it as a gap in problem understanding, where insufficiently grounded agents cannot move beyond canonical methods, whereas the DeepSurvey-condition agent could attempt a genuinely novel design and attain the best outcome within the same session budget.

D Main Evaluation Supplement

D.1 Test Dataset

We will present the specific input topics in our evaluation setup in this section:

- **Biology**
  - Microbial Enzymes in Pollutant Bioremediation
  - Lactic Acid Bacteria and Bacteriocins
  - Mathematical Models of Malaria
  - Single-cell RNA Sequencing and Spatial Transcriptomics
  - Multi-omics Integration for Biological Discovery
- **Medicine**
  - Dietary Sugars and Body Weight in Randomised Controlled Trials
  - Flavonoids and Phenolic Compounds from Medicinal Plants
  - Patient Engagement in Research
  - Wearable Health Monitoring
  - Gut Microbiome and Disease
- **Psychology**
  - Neuroimaging Studies of Internet and Gaming Addiction
  - Technology-Delivered Interventions for Youth Depression and Anxiety
  - Social Media and Mental Health
  - Trauma and PTSD Intervention
  - Digital Mental Health
- **Computer Science**
  - Multimodal Machine Learning Taxonomy
  - LLMs-based Agent
  - 3D Gaussian Splatting
  - Hallucination in LLMs
  - Acceleration for LLMs
- **Materials Science**
  - Solid-State Batteries
  - Perovskite Solar Cells
  - Two-Dimensional Materials and Their Applications in Electronics and Energy
  - Metal-Organic Frameworks for Energy and Environmental Applications
  - Advanced Battery Materials for Next-Generation Energy Storage

Topics are sourced from SurveyGen's topics with high citation count and some popular topics from Semantic Scholar, balancing influence and recency.

D.2 Detailed Evaluation Criteria

This section provides detailed definitions for the content quality sub-dimensions described in the main paper (Section 3.1.1). Our evaluation adopts a hierarchical scoring framework where content quality is decomposed into three aggregated dimensions, each comprising fine-grained sub-dimensions scored on a 1–10 scale.

Content Score aggregation. Each aggregated dimension is computed as the arithmetic mean of its sub-dimension scores:

$$\text{Core} = \frac{1}{4}(\text{Synthesis} + \text{Organization} + \text{Comprehensiveness} + \text{Relevance}) \tag{7}$$

$$\text{Writing} = \frac{1}{3}(\text{Readability} + \text{Academic Rigor} + \text{Clarity \& Coherence}) \tag{8}$$

$$\text{Depth} = \frac{1}{4}(\text{Critical Analysis} + \text{Novelty and Insights} + \text{Specificity} + \text{Future Directions}) \tag{9}$$

The final content quality score combines the three dimensions with weights reflecting our definition of a high-quality survey:

$$\text{Total} = 0.4 \times \text{Core} + 0.2 \times \text{Writing} + 0.4 \times \text{Depth} \tag{10}$$

Each generated survey is independently scored on every sub-dimension, with the average across multiple evaluation rounds reported as the final result to improve stability.

Since a self-designed criterion would probably introduce preference toward DeepSurvey, our evaluation formula is built on the existing Agentic AutoSurvey's protocol[6] (Liu et al., 2025b) for reasonable and objective evaluation. The final weights were derived by adjusting Agentic AutoSurvey's formula according to our core motivation. Agentic AutoSurvey uses weights of 0.6/0.2/0.2 for Core Quality/Writing Quality/Content Depth. We consider that this formula overemphasizes Core Quality which primarily reflects top-level requirements (like relevance and comprehensiveness) that existing methods (e.g., AutoSurvey (Wang et al., 2024b)) already handle reasonably well. Moreover, as surveys in our setting serve as knowledge sources for downstream research agents, the breadth reflected in Core Quality provides relatively limited additional value to such consumers; the representation emphasized by Writing Quality is similarly of secondary importance. To reflect the premise that analytical depth, rather than breadth or representation, constitutes the principal knowledge a survey adds for such consumers, we moderately reduced the Core Quality weight and increased the Content Depth weight, yielding the final 0.4/0.2/0.4 weight.

Weight Robustness. A potential concern is that our re-weighted formula could bias the comparison in favor of DeepSurvey. To rule this out, we recompute the total score of all methods directly from the fine-grained dimension scores in Table 1 under two alternative weight configurations: Agentic AutoSurvey's original formula (0.6/0.2/0.2) and uniform weights (1/3 each). As shown in Table 8, DeepSurvey remains the top method under all three configurations, outperforming the strongest baseline SurveyForge by comparable margins.

Table 8: Total content scores under three weight configurations over Core Quality/Writing Quality/-Content Depth, computed from the fine-grained scores in Table 1. The ranking is unchanged across configurations, confirming that DeepSurvey's advantage does not depend on our choice of weights.

Method	Ours (0.4/0.2/0.4)	Agentic AutoSurvey (0.6/0.2/0.2)	Uniform (1/3, 1/3, 1/3)
DeepSurvey	**8.341**	**8.535**	**8.315**
SurveyForge	8.071	8.297	8.085
AutoSurvey	7.958	8.252	7.990
SurveyX	6.993	7.478	7.006
LiRA	5.687	5.544	5.873

Furthermore, in Section 3.6, we validated the consistency of our total score formula with human expert preferences. We believe this meta-evaluation substantiates the rationality of our total score formula.

[6]Since the code of Agentic AutoSurvey has not been open-sourced, we could not include it as a baseline in our main experiments.

Citation Quality. Citation quality evaluates the accuracy, relevance, and completeness of references in generated surveys. We adopt NLI-based metrics following AutoSurvey (Wang et al., 2024b), which have been widely used in subsequent work (Wang et al., 2025; Liang et al., 2025; Bao et al., 2025). The core idea is to judge whether each cited reference actually supports the corresponding claim in the survey.

Specifically, the evaluation first extracts all citation-bearing claims $C = \{c_1, c_2, \ldots\}$ from the survey, where each claim c_i is associated with a set of cited references $\text{Ref}_i = \{r_{i1}, r_{i2}, \ldots\}$. An NLI model h then determines whether the reference set entails the claim.

Citation Recall measures whether every claim in the survey is supported by its cited references—i.e., the completeness of evidential coverage:

$$R = \frac{\sum_{i=1}^{|C|} \mathbb{I}[h(c_i, \text{Ref}_i) = 1]}{|C|} \tag{11}$$

where $h(c_i, \text{Ref}_i) = 1$ indicates that the reference set collectively supports claim c_i, $\mathbb{I}[\cdot]$ is an indicator function, whose value is 1 when the condition is true, and 0 otherwise. A high recall means the survey's claims are well-grounded in cited literature.

Citation Precision measures whether each individual reference is actually relevant to the claim it supports—i.e., the accuracy of citation assignment. We define an auxiliary function g to assess the necessity of each reference:

$$g(c_i, r_{ik}) = \mathbb{I}[h(c_i, \{r_{ik}\}) = 1] \vee \mathbb{I}[h(c_i, \text{Ref}_i \setminus \{r_{ik}\}) = 0] \tag{12}$$

This function judges that reference r_{ik} is relevant to claim c_i if either (1) r_{ik} alone can support c_i, or (2) removing r_{ik} from Ref_i causes the remaining set to fail to support c_i (i.e., r_{ik} is necessary). Precision is then computed as:

$$P = \frac{\sum_{i=1}^{|C|} \sum_{k=1}^{|\text{Ref}_i|} \mathbb{I}[h(c_i, \text{Ref}_i) = 1 \wedge g(c_i, r_{ik}) = 1]}{\sum_{i=1}^{|C|} |\text{Ref}_i|} \tag{13}$$

A high precision means that cited references are genuinely relevant rather than spurious.

Valid Citation Ratio reports the fraction of citations that are successfully parsed and used for evaluation, excluding entries with formatting errors or unresolvable references:

$$\text{Valid Ratio} = \frac{|\text{Valid Citations}|}{|\text{Total Citations}|} \tag{14}$$

This metric serves as a basic sanity check: a ratio below 1.0 indicates citation formatting issues or hallucinated references that could not be matched to any paper in the database.

The full evaluation pipeline proceeds as follows: (1) extract citation-bearing claims from the survey text and identify their reference numbers; (2) use the NLI model to judge entailment between each claim and its reference set; (3) for supported claims, further assess the relevance of individual references via the auxiliary function g; (4) aggregate all metrics into the final citation quality report.

D.3 Detailed Evaluation Standard

The complete 1–10 level descriptors for every sub-dimension, including the two supplementary code evaluation dimensions for CS surveys, are presented in Tables 29–32 at the end of the appendix.

Agent-Oriented Adaptation of the Protocol. Beyond re-weighting the three dimensions, we calibrate several sub-dimensions toward the needs of agent consumers, following the premise in Section 1 that agent readers already hold broad knowledge and require analytical depth and reliable attribution more than restated breadth. The adaptations are as follows.

Comprehensiveness is calibrated to measure whether the survey covers the information needed for further research within the topical scope, such as representative methods, key mechanisms, and standard experimental settings, rather than the exhaustive inclusion of tangential material. For an agent, missing research-critical material is costly, whereas restating well-known background adds little.

Future Directions is calibrated to reward concrete and actionable directions, including specific open problems, candidate approaches, and implementation-oriented analysis, rather than generic outlook statements. This measures whether the survey can directly inform an agent's follow-up experiment design and idea generation.

Specificity is introduced as an additional sub-dimension beyond Agentic AutoSurvey's protocol. Under this premise, the information a survey adds for agent consumers concentrates in fine-grained mechanisms, design choices, and experimental configurations rather than in high-level summaries that largely restate established work; a granularity-sensitive dimension therefore measures how well a survey supplies exactly the knowledge that downstream agents cannot recover from parametric knowledge alone.

Critical Analysis emphasizes the trade-offs, applicability boundaries, and failure modes of existing methods, which an agent needs to avoid redundant or already-refuted research directions.

Writing Quality is retained as a sanity check for human compatibility but down-weighted: agents consume survey content as evidence rather than as prose, so readability affects their consumption far less than it affects human readers. Within this dimension, we further merge the separate Clarity and Coherence sub-dimensions of Agentic AutoSurvey into a single Clarity & Coherence criterion, which raises the relative weight of Academic Rigor; this reflects that Academic Rigor, which scores citation practice and the fair representation of prior work, is the content-side counterpart of citation reliability and thus carries greater significance for agent consumers than purely presentational aspects.

D.4 Detailed Evaluation Parameters

Table 23 lists the key system parameters used in DeepSurvey. Note that we use MiniLMv2 (Wang et al., 2020) as embedding model. For DeepSurvey model, we choose MiMo-V2.5 for its lower cost and fast speed. In addition, MiMo-V2.5 is an open-source model that facilitates reproduction. The primary judge model is MiMo-V2.5[7], used for both content scoring and the NLI-based citation evaluation. For robustness, we additionally evaluate with judge models from other families, including GPT-5-mini (Singh et al., 2026), Gemini-3[8], DeepSeek-V3 (DeepSeek-AI et al., 2024), GPT-5.4, and MiMo-V2-Flash (Xiao et al., 2026) (Appendix D.9).

D.5 Stability and Statistical Significance

Tables 24, 25, 26, 27, and 19 report the mean $\pm$ standard deviation (with coefficient of variation, CV) of each metric over repeated runs, and Table 9 summarizes the mean differences and effect sizes (Hedges' g) of DeepSurvey against each baseline.

Score Stability. DeepSurvey's aggregated scores are highly stable, with CVs of 0.8%–2.2% across the four dimensions—the lowest worst-case fluctuation among all methods—and its lead over the strongest baseline (e.g., $+0.27$ in Total Score over SurveyForge) far exceeds this noise level. Fine-grained dimensions remain similarly stable (CVs of 1.0%–6.3%), while the largest fluctuations occur in sub-dimensions of weaker baselines (e.g., Comprehensiveness for LiRA, 10.3%; Academic Rigor for SurveyX, 10.6%). Citation metrics fluctuate more, as they are computed at the claim level (CVs of 8.7% and 7.5% for DeepSurvey's precision and recall), yet stay within the range of the baselines (up to 9.9%).

Statistical Significance. All comparisons reach a large effect ($g \geq 0.8$, Cohen's convention) except Writing Quality against AutoSurvey and SurveyForge (both $g = 0.11$), and Core Quality against SurveyForge ($g = 0.29$). In particular, the advantages in citation reliability (Precision: $g = 2.08$–5.90; Recall: $g = 2.06$–8.19), Content Depth ($g = 0.87$–5.35), and Total Score ($g = 1.25$–8.11) are uniformly large, confirming that the depth and reliability gains are statistically meaningful rather than run-to-run noise. The only weak effects concentrate in Writing Quality, where DeepSurvey is closest to the strongest baselines, consistent with the main-text results.

[7] https://mimo.xiaomi.com/mimo-v2-5

[8] https://blog.google/products-and-platforms/products/gemini/gemini-3

Table 9: Statistical significance of DeepSurvey vs. each baseline. Each cell reports the mean difference (left of "/") and the effect size (Hedges' g, right of "/"). Bold g values (≥ 0.8) indicate a large effect by Cohen's convention; non-bolded values are small ($0.2 \leq g < 0.5$) or negligible ($g < 0.2$).

Metric	vs AutoSurvey	vs SurveyForge	vs SurveyX	vs LiRA
Citation Quality				
Precision	0.23 / **3.08**	0.15 / **2.08**	0.44 / **5.90**	0.34 / **4.98**
Recall	0.19 / **4.97**	0.13 / **2.06**	0.44 / **8.19**	0.19 / **6.37**
Content Quality				
Total Score	0.38 / **1.46**	0.27 / **1.25**	1.35 / **4.76**	2.65 / **8.11**
Core Quality	0.22 / **1.06**	0.25 / 0.29	0.68 / **3.11**	3.82 / **6.96**
Writing Quality	0.03 / 0.11	0.03 / 0.11	1.12 / **2.63**	1.38 / **7.52**
Content Depth	0.73 / **1.55**	0.41 / **0.87**	2.13 / **5.35**	2.13 / **5.11**

D.6 Cost Evaluation

Time&Token Cost We run DeepSurvey (under the setting in Appendix D.4) and baseline methods over multiple runs and calculate the average cost using the API provider's pricing information. Results are shown in Table 10.

Table 10: Cost comparison across methods.

Method	Time	Token	Money-Cost
DeepSurvey	1407s	2.28M	\$0.474
AutoSurvey	844s	2.86M	\$0.372
SurveyForge	902s	4.58M	\$0.794
LiRA	473s	0.31M	\$0.060
SurveyX	3873s	3.42M	\$0.614

As can be seen, apart from LiRA's significantly lower costs, DeepSurvey's costs are comparable to other baselines—higher than AutoSurvey but lower than SurveyX and SurveyForge. In terms of time cost, DeepSurvey is higher than AutoSurvey and SurveyForge and lower than SurveyX.

Though DeepSurvey does not have the lowest cost, we consider this acceptable in exchange for surveys with greater analytical depth and citation reliability, which is our primary motivation.

Table 11: Paper budget across domains. Note: Bio=Biology, Med=Medicine, Psy=Psychology, CS=Computer Science, MatSci=Materials Science.

Metric	Bio	Med	Psy	CS	MatSci	Overall
citation/subsection	7.544	8.201	8.927	9.592	7.312	8.315
citation/section	20.805	21.582	24.667	26.320	16.699	22.015
citation/paper	63.600	61.800	62.800	105.200	44.600	67.600

Paper Budget. As paper budgets may vary significantly across domains, we supplement additional experiments across 5 domains in Table 11. Across all domains, the average paper budget per survey is 67.6, and 105.2 in Computer Science. The average per-section budget is 22.015. These results are averaged over multiple random runs.

Relationship of Cost and Paper Budget We experiment on the relationship between time&token cost and the paper budget for further exploration. Results are shown in Figure 8.

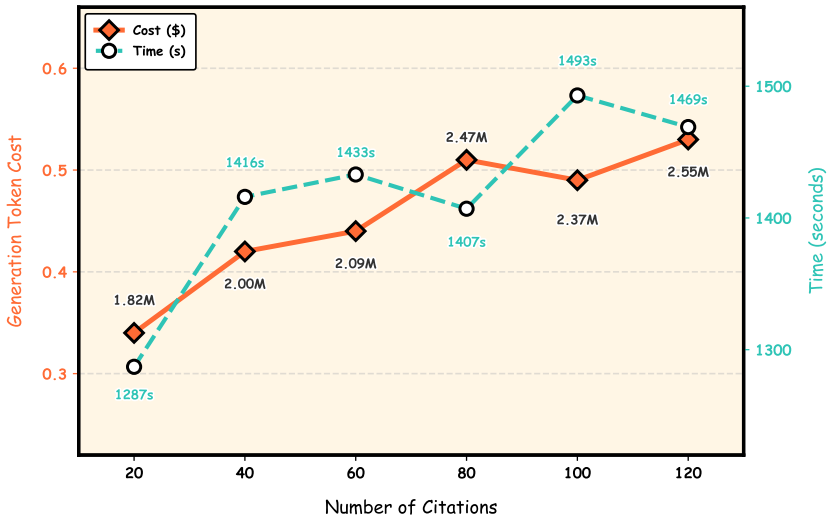


Figure 8: Generation cost, runtime, and token consumption as functions of the number of citations. The left y-axis shows the generation cost in USD, the right y-axis shows the runtime in seconds, and the token usage (in millions) is annotated above each data point.

Table 12: Linear Scaling Relationship between Citation Budget and Generation Cost.

Metric	Pearson r	p-value	R^2	Slope (per citation)
Time (s)	0.8307	0.0406	0.6901	1.5929
Cost ($)	0.9383	0.0056	0.8804	0.0018
Token (M)	0.9486	0.0039	0.8999	0.0073

We briefly analyze the linear scaling in Table 12. Cost and tokens scale roughly linearly with the citation count (Pearson r > 0.9), while time shows a much weaker linear relationship. We attribute this to optimizations such as parallelization and caching, which prevent time from growing linearly with the number of citations.

D.7 Human Expert Experiment Supplement

Table 13: Overview of the 6 Evaluation Topics.

No.	Paper Title	Publication Date	Citations
1	Reinforcement Learning for Large Reasoning Models	2025 Sep	158
2	Agentic Reinforcement Learning for LLMs	2025 Sep	163
3	Agentic AI Autonomous Intelligence for Complex Goals	2025	576
4	Multi-Functional and Hybrid Reconfigurable Intelligent Surfaces for Integrated Sensing and Communication	2025 Oct	70
5	Vision-Language-Action Models: An Action Tokenization Perspective	2025 Jul	93
6	Edge General Intelligence with Agentic AI and Agentification	2025 Aug	56

Detailed Report on Human Expert Evaluation The evaluation set contains 6 influential and popular topics, summarized in Table 13. The source papers and metadata are obtained from Semantic Scholar. Each topic contains one human-written survey and one DeepSurvey-generated survey. Each topic is annotated by 4 annotators. All annotations are performed in a blinded setting. The final win rates are aggregated over all annotation samples.

We have excluded the corresponding human-written survey for each topic from the paper pool during DeepSurvey's generation process to prevent data leakage.

The evaluators were recruited from students and researchers with academic writings and evaluation experience in corresponding scientific domains. Participation was voluntary. No crowdworkers were used.

Requirements for Human Expert

Annotation Requirements

```
Annotation Task Instructions

You will complete multiple comparison groups. Each task contains
    two surveys, labeled A and B. Please read both surveys in each
    group, and then judge which is better on the following three
    dimensions, and give an overall judgment. The data obtained
    from this task will be used exclusively for research purposes.
    Any results made public will be reported only in aggregate, and
    no identifying information (such as your name or other personal
    identifiers) will be included.

Annotation Format
If A is better than B on a dimension, fill in A; if B is better
    than A, fill in B; if they are tied, fill in Tie.
Example: If A is better than B on Core Quality, note it as Core
    Quality: A.

Group 1
Core Quality:
Writing Quality:
Content Depth:
Overall Judgment:

Group 2
Core Quality:
Writing Quality:
Content Depth:
Overall Judgment:

...

Brief Description of Annotation Dimensions

1. Core Quality
Measures whether the survey stays closely on topic, effectively
    integrates the literature, and builds a systematic synthesis.
Key aspects: synthesis, organization, comprehensiveness, relevance.
Staying on-topic and building a systematic synthesis.

Synthesis: integrates papers into a coherent whole, not just listing
Organization: logical section flow and hierarchy
Comprehensiveness: covers key methods, works, branches
Relevance: minimal off-topic content

2. Writing Quality
Measures whether the language expression is clear, rigorous, and
    coherent, and whether technical details are accurately conveyed.
Key aspects: readability, academic rigor, clarity & coherence.
Clear, rigorous, and coherent writing.

Readability: fluent, natural, easy to follow
Academic Rigor: precise, well-supported academic expression
Clarity & Coherence: clear technical descriptions; consistent
    narrative across sections
```

```
3. Content Depth
Measures whether the survey goes beyond simple summary to offer
    critical analysis, unique insights, and actionable research
    suggestions.
Key aspects: critical analysis, novelty & insights, specificity,
    future directions.
Going beyond summary to offer insight and actionable directions.

Critical Analysis: discusses strengths, limitations, boundaries
Novelty & Insights: new perspectives or cross-work understanding
Specificity: fine-grained technical/experimental details
Future Directions: concrete, feasible, gap-linked suggestions

Overall Judgment
Please synthesize the performance across the above three dimensions
    and provide an overall comparison result for the two reviews.
```

D.8 Citation Density Analysis

A natural question is whether the citation reliability gains in Table 1 merely reflect differences in citation volume, since both recall and precision are normalized over claims and references. We therefore compute citation density (citations per sentence) for DeepSurvey and the two strongest baselines, aggregated over all surveys generated in the main experiment. Micro density is computed as total citations divided by total sentences; macro density is the mean of per-survey densities.

Table 14: Citation density comparison. "Micro" divides total citations by total sentences; "Macro" averages per-survey densities.

Method	Total Sentences	Total Citations	Density (micro)	Density (macro)
DeepSurvey	7,335	2,105	0.287	0.289
AutoSurvey	7,931	1,972	0.249	0.254
SurveyForge	6,670	2,051	0.307	0.303

Three observations follow. First, DeepSurvey issues the largest number of citations (2,105) while keeping a mid-range density (0.287 micro / 0.289 macro), between AutoSurvey (0.249 / 0.254) and SurveyForge (0.307 / 0.303): it neither inflates density, which would risk precision by attaching loosely related references, nor reduces it, which would sacrifice recall by leaving claims unsupported. In addition, the recall gap between DeepSurvey (0.630) and SurveyForge (0.503) cannot be explained by citing more per sentence: SurveyForge cites more often per sentence than DeepSurvey yet supports a smaller fraction of its claims, indicating that assignment quality, not volume, drives recall.

D.9 Evaluation with Other Model

Since different LLMs differ in judging, adding experiments with judges from other families makes the results more convincing. We supplement evaluations with GPT-5-mini, Gemini-3-Flash and MiMo-V2-Flash as judge models (Tables 15, 16, and 17).

Although these models show different scoring preferences, DeepSurvey consistently and substantially outperforms the strongest baseline across all three judges (7.982 vs. 7.633, 8.887 vs. 8.158, 8.676 vs. 8.483). Our conclusions remain unchanged under different judge model families.

Table 15: Detailed evaluation results with GPT-5-mini.

Method	Total_Score	Core_Quality	Writing_Quality	Content_Depth	Precision	Recall
DeepSurvey	**7.982**	**8.483**	7.412	**7.767**	**0.560**	**0.618**
AutoSurvey	7.633	8.354	**7.500**	6.979	0.485	0.571
LiRA	7.236	7.659	7.364	6.750	0.129	0.290
SurveyForge	7.529	8.396	6.979	6.938	0.420	0.462
SurveyX	5.887	6.633	5.533	5.317	0.276	0.312

Table 16: Detailed evaluation results with Gemini-3-Flash.

Method	Total_Score	Core_Quality	Writing_Quality	Content_Depth	Precision	Recall
DeepSurvey	**8.887**	**9.300**	**8.533**	**8.650**	**0.475**	**0.541**
AutoSurvey	6.967	7.500	7.583	6.125	0.445	0.489
LiRA	7.344	7.667	7.556	6.917	0.066	0.224
SurveyForge	8.158	8.875	7.667	7.688	0.273	0.312
SurveyX	5.747	6.950	5.933	4.450	0.154	0.187

Table 17: Detailed evaluation results with MiMo-V2-Flash.

Method	Total_Score	Core_Quality	Writing_Quality	Content_Depth	Precision	Recall
DeepSurvey	**8.676**	**9.083**	8.311	**8.450**	**0.700**	**0.654**
AutoSurvey	8.483	8.938	**8.417**	8.063	0.648	0.623
SurveyForge	8.101	8.708	7.795	7.646	0.379	0.395
SurveyX	7.244	7.667	7.222	6.833	0.344	0.287
LiRA	8.004	8.455	7.793	7.659	0.337	0.211

E Ablation Experiment Supplement

E.1 Multi-granularity Refinement Component Ablation

The iterative refinement module operates at three granularity levels: section-level, subsection-level, and survey-level. To assess the individual contribution of each, we ablate them independently. w/o Section Refinement removes section-level optimization; w/o Subsection Refinement removes subsection-level optimization; w/o Survey Refinement removes survey-level optimization; w/o All removes all three. Note that the base model for this experiment is MiMo-V2-Flash rather than MiMo-V2.5, so the data differs from the main text.

Table 18: Ablation results on refinement components.

Method	Total	Core	Writing	Depth
DeepSurvey	**8.676**	9.083	8.311	**8.450**
w/o Sec.	8.544	9.042	**8.389**	8.125
w/o Sub.	8.487	9.000	8.333	8.050
w/o Sur.	8.487	9.050	8.333	8.000
w/o All	8.348	**9.107**	7.810	7.857

Table 18 presents the main dimension results and Table 22 presents the detailed dimension results. We highlight three key findings.

(1) **Subsection-level refinement is a critical component.** Removing it causes a substantial total score drop (8.676→8.487, Δ=0.189) and a substantial content depth decrease (8.450→8.050, Δ=0.400). Since subsections constitute the basic writing unit of a survey, optimizing at this granularity directly improves the quality of individual research discussions, yielding the strongest impact on overall performance.

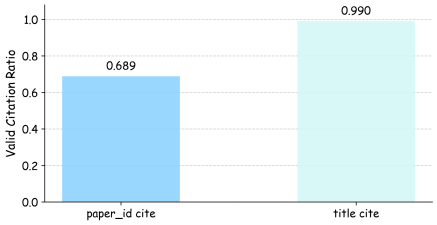


Figure 9: Valid citation ratio comparison between paper ID and paper title citation marks.

(2) **Survey-level refinement is indispensable for cross-section consistency.** Although its total score impact equals that of subsection-level removal (both Δ=0.189), survey-level refinement uniquely affects content depth more severely (Δ=0.450, the largest single-component depth drop). This indicates its irreplaceable role in harmonizing cross-section terminology, eliminating redundancy, and ensuring global narrative coherence—issues that local refinement cannot address.

(3) **The three levels are complementary, not redundant.** Removing all refinement (w/o All) degrades writing quality to 7.810 and content depth to 7.857, both significantly worse than any single ablation. This confirms that subsection-level refinement targets local readability, section-level refinement enhances analytical depth within sections, and survey-level refinement maintains global consistency. Together they form a complete multi-granularity optimization system.

From a resource allocation perspective, subsection-level and survey-level refinement each contribute 0.189 to the total score, while section-level contributes 0.132. When computational budget is constrained, prioritizing subsection and survey-level refinement yields the highest marginal benefit.

E.2 Citation Mark Ablation

DeepSurvey uses paper titles (e.g., `<AutoSurvey: Large Language Models Can Automatically Write Surveys>`) as citation marks in generated surveys. To validate this design choice, we compare it against an alternative using paper IDs (e.g., `<2406.10252>`). The two strategies represent a trade-off: paper IDs are short and format-uniform, reducing context window consumption and enabling direct lookup via academic databases, whereas paper titles carry richer semantic information that may help the LLM more accurately associate claims with their supporting references, but at the cost of significantly increased input length.

Figure 9 shows that title-based citation marks achieve a substantially higher valid citation ratio. A plausible explanation is that paper titles provide explicit semantic cues—such as method names, task descriptions, and key terms—that help the LLM ground its citations during generation, reducing hallucinated or misattributed references. In contrast, paper IDs are semantically opaque identifiers that require the model to maintain ID–content mappings from context alone, a burden that intensifies in long-text generation scenarios. This finding suggests that the representational format of input information materially affects LLM generation quality, and that careful design of citation representations can improve reliability without modifying the underlying model.

E.3 Draft Quality Improvement in Refinement

To validate that the multi-granularity refinement subsystem (Section 2.5) genuinely improves draft quality rather than merely altering surface expression, we conduct a controlled evaluation comparing survey drafts before and after refinement. We evaluate whether this multi-granularity coordination yields measurable quality gains.

Specifically, we employ a panel of expert LLMs spanning multiple model families (GPT-series, Gemini-series, and DeepSeek-series) to independently assess draft quality along four dimensions before and after refinement: **critical analysis depth** (whether arguments are substantiated with comparative reasoning rather than superficial description), **logical coherence** (whether paragraphs follow a clear argumentative thread with appropriate transitions), **academic rigor** (whether claims are properly qualified and supported by citations), and **readability** (whether the text is clear, concise,

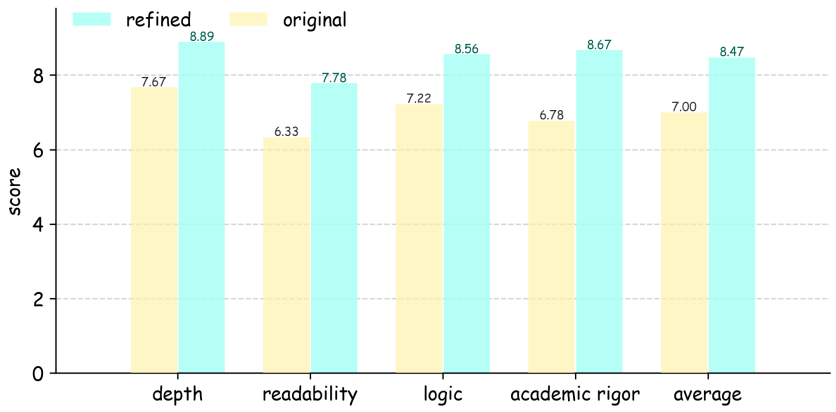


Figure 10: Quality dimension scores before and after multi-granularity refinement, averaged across expert LLM judges.

and free of redundancy). Each model scores drafts on a 1–10 scale for each dimension, and we report the average across all models.

Figure 10 presents the results. The refinement subsystem yields consistent improvements across all four dimensions, with the most pronounced gains in critical analysis depth and academic rigor. This aligns with the design intent: the reviewer role specifically targets analytical weaknesses and citation issues, while the multi-granularity optimization strategy addresses coherence at local (subsection), structural (section), and global (survey) levels.

E.4 Detailed Scores of the Module Ablation

Due to space constraints in the main body, we present fine-grained dimension scores for the module ablation in Table 21, highlighting key findings that complement the aggregated analysis in Section 3.5.

The collector is the primary bottleneck. Removing graph-based collection causes catastrophic drops in Comprehensiveness (8.889→3.333) and Relevance (9.300→1.333), confirming that without curated evidence, the system cannot cover the topic space. The damage extends to Academic Rigor (8.506→6.000), Clarity & Coherence (8.333→4.333), and several depth metrics (e.g., Specificity 8.194→6.942), indicating that evidence scarcity degrades not only coverage but also the rigor and coherence of the generated text.

The writer primarily governs synthesis, organization, and academic rigor. Its removal causes Synthesis Quality (8.334→7.667) and Organization (8.935→7.333) to drop sharply and collapses Academic Rigor (8.506→5.333) and Critical Analysis (7.389→5.341), both the lowest among all ablations, while Comprehensiveness (8.000) and Relevance (8.933) are comparatively less affected. This confirms that outline-driven writing with explicit citation assignment is essential for well-grounded, insightful narratives rather than for evidence coverage per se.

The analyzer contributes primarily to depth-oriented metrics. Its removal most affects Specificity (8.194→7.286, the largest single-metric drop for this component) and Novelty and Insights (7.866→7.111), while Comprehensiveness and Relevance remain near ceiling (8.857 and 9.000), consistent with its role in constructing relation graphs and comparison tables for cross-paper synthesis.

Refinement polishes rather than transforms. Removing refinement causes moderate declines across most metrics rather than catastrophic failures, with the largest drops in Novelty and Insights (7.866→6.948), Organization (8.935→8.123), Clarity & Coherence (8.333→7.528), and Future Directions (8.139→7.333), consistent with its role in polishing insight expression and global coherence.

The remaining variants cause milder, depth-oriented declines. w/o keynote, w/o agentic, and w/o code all reduce Novelty and Insights (to 6.667, 7.635, and 7.533) and Specificity (to 7.671, 7.914, and 7.833), consistent with the aggregated results in Section 3.5, where each incurs its largest drop in Content Depth.

F Analysis Result Cases

In this section, we demonstrate specific cases from the global knowledge substrate.

F.1 Keynote Case

We use 'AutoSurvey' as the input topic. The generated keynotes are as follows:

Keynotes Case


```
"title": "AutoSurvey: Large Language Models Can Automatically Write
    Surveys",
"authors": ["Yidong Wang", "Qi Guo", "Wenjin Yao", "Hongbo Zhang",
    "Xin Zhang", "Zhen Wu", "Meishan Zhang", "Xinyu Dai", "Min
    Zhang", "Qingsong Wen", "Wei Ye", "Shikun Zhang", "Yue Zhang"],
"abstract": "This paper introduces AutoSurvey, a speedy and
    well-organized methodology for automating the creation of
    comprehensive literature surveys in rapidly evolving fields
    like artificial intelligence. Traditional survey paper creation
    faces challenges due to the vast volume and complexity of
    information, prompting the need for efficient survey methods.
    While large language models (LLMs) offer promise in automating
    this process, challenges such as context window limitations,
    parametric knowledge constraints, and the lack of evaluation
    benchmarks remain. AutoSurvey addresses these challenges
    through a systematic approach that involves initial retrieval
    and outline generation, subsection drafting by specialized
    LLMs, integration and refinement, and rigorous evaluation and
    iteration.",
"key_contributions": [
    "Introduction of AutoSurvey, a comprehensive system for
    automatically generating academic surveys using LLMs.",
    "A two-stage parallel generation approach (outline generation
    and subsection drafting) to overcome context window limitations
    and speed up the process.",
    "Integration of Retrieval-Augmented Generation (RAG) for
    real-time knowledge updates to ensure citations are current and
    accurate."
    "Development of a Multi-LLM-as-Judge evaluation framework with
    metrics for citation quality (recall and precision) and content
    quality (coverage, structure, relevance).",
    "Experimental validation demonstrating AutoSurvey\'s
    effectiveness in terms of speed, citation quality, and content
    quality, approaching human performance.",
    "Open-sourcing of resources and prompts to facilitate further
    research."],
"methodology": "AutoSurvey follows a four-phase pipeline:
    (1) Initial Retrieval and Outline Generation, where a database
    of papers is scanned using embedding-based retrieval to
    generate a structured outline;
    (2) Subsection Drafting, where specialized LLMs draft each
    section in parallel using the outline and retrieved papers;
    (3) Integration and Refinement, where sections are refined for
    coherence, corrected for citation errors, and merged into a
    cohesive document;
```

```
    (4) Rigorous Evaluation and Iteration, where the survey is
    assessed using Multi-LLM-as-Judge and the best among multiple
    candidate surveys is selected. The system uses a retrieval
    database of 530,000 arXiv papers and leverages parallel
    processing to enhance efficiency."
"experiments": {
    "setup": "Comparisons were made against human-authored surveys
    and naive RAG-based LLM generation across 20 computer science
    topics. Surveys of varying lengths (8k, 16k, 32k, 64k tokens)
    were generated. Claude-3-Haiku was used as the base writer for
    AutoSurvey and naive RAG, while evaluations employed a
    combination of GPT-4, Claude-3-Haiku, and Gemini-1.5-Pro.
    Metrics included survey creation speed, citation quality
    (recall and precision), and content quality (coverage,
    structure, relevance) scored on a 5-point rubric.",
    "baselines": [
        "Human-authored surveys collected from arXiv",
        "Naive RAG-based LLM generation (iterative prompting with
    retrieved papers)"],
    "main_results": {
        "speed": "AutoSurvey is significantly faster: for 64k-token
    surveys, it generates 73.59 surveys per hour vs. 12.56 for
    naive RAG and 0.07 for humans.",
        "citation_quality": "AutoSurvey achieves near-human recall
    and precision (e.g., at 64k tokens: recall 82.25%, precision
    77.41% vs. human 86.33% and 77.78%), outperforming naive RAG
    (68.79% and 61.97%).",
        "content_quality": "AutoSurvey scores highly (e.g., at 16k
    tokens: coverage 4.66, structure 4.33, relevance 4.86, average
    4.60) compared to human (4.66, 4.38, 5.00, 4.66) and naive RAG
    (4.46, 3.66, 4.73, 4.23)."},
        "additional_studies": {
            "meta_evaluation": "Spearman\'s rank correlation
    between LLM-based and human rankings showed moderate to strong
    positive correlation (0.5429 for mixed models), indicating
    alignment with human preferences.",
            "ablation_study": "Removing retrieval drastically
    reduces citation quality; removing reflection slightly impacts
    content quality. Different base LLMs (GPT-4, Claude-haiku,
    Gemini-1.5-Pro) all yield competitive results, with GPT-4
    slightly best.",
            "knowledge_enrichment_test": "Providing
    AutoSurvey-generated surveys improved accuracy on
    multiple-choice questions by 9.2% over direct answers and 2.4%
    over naive RAG surveys."}},
"results": "AutoSurvey demonstrates a strong balance between speed
    and quality, achieving near-human performance in citation and
    content quality while being orders of magnitude faster. It
    outperforms naive RAG-based generation in all metrics,
    particularly in structure and citation accuracy. The system is
    robust across different survey lengths and base LLMs.",
"significance": "AutoSurvey addresses a critical bottleneck in
    academic research: the time-consuming process of literature
    review. By automating survey generation with high quality and
    speed, it can accelerate knowledge synthesis, help researchers
    stay current in fast-moving fields, and fill gaps in survey
    coverage. The open-sourced methodology and evaluation framework
    provide a foundation for future work in automated academic
    writing.",
"limitations": [
    "Citation errors, particularly overgeneralization (51% of
    errors), indicate reliance on parametric knowledge over
    retrieved information.",
```

```
      "The evaluation, while correlated with human judgment, is still
      automated and may not capture all nuances.",
      "The system is currently focused on computer science topics;
      generalizability to other domains needs exploration.",
      "Ethical considerations around authorship and potential misuse
      of auto-generated surveys."],
      "future_directions": [
      "Improving citation accuracy by reducing overgeneralization
      through better integration of retrieved knowledge.",
      "Extending the system to other domains and document types
      (e.g., project reports, white papers).",
      "Enhancing the evaluation framework with more nuanced metrics
      and human-in-the-loop refinement.",
      "Exploring iterative refinement where the system incorporates
      feedback from initial drafts to improve subsequent versions.",
      "Investigating the use of more advanced or specialized LLMs to
      further boost quality."],
  "critical_reflections": [
      "AutoSurvey\'s parallel generation approach elegantly tackles
      context window limits, but the merging step may introduce
      coherence issues; the refinement phase mitigates this but adds
      computational cost.",
      "The Multi-LLM-as-Judge strategy reduces single-model bias, yet
      the moderate correlation with human experts suggests room for
      improvement in evaluation fidelity.",
      "While AutoSurvey outperforms naive RAG, the gap between it and
      human performance in structure and overall quality indicates
      that LLMs still struggle with holistic document organization
      and logical flow.",
      "The cost analysis shows AutoSurvey is cost-effective (e.g.,
      \$1.2 per 32k survey with Claude-haiku), making it practical
      for large-scale use.",
      "The system\'s reliance on arXiv papers may limit coverage of
      non-open-access literature, potentially introducing source
      bias.",
      "The example survey on emotion recognition (Appendix F)
      demonstrates the system\'s capability to generate coherent
      outlines and content, but human oversight is still recommended
      for final quality assurance."],
      "tldr": "AutoSurvey automates the creation of academic surveys
      by leveraging LLMs in a parallel, retrieval-augmented pipeline:
      it generates outlines, drafts sections concurrently, refines
      for coherence, and evaluates with multi-LLM judges. It
      significantly speeds up survey writing while achieving citation
      and content quality close to human experts, outperforming naive
      RAG approaches. This offers a scalable tool for synthesizing
      research literature in fast-evolving fields."
```


F.2 Clustering Result

We use 'AutoSurvey' as the input topic. The clustering result is shown in Table 28.

F.3 Intra-cluster Q&A Analysis Example

We use "AutoSurvey" as the input topic for retrieval, clustering, and intra-cluster Q&A analysis. One example of the Q&A analysis is shown below.

Cluster 1 Question: Several papers propose explicit hierarchical structures (e.g., knowledge trees, citation graphs) to organize surveys, while others rely on iterative outline refinement without such explicit models. How do these fundamentally different approaches to structural planning compare

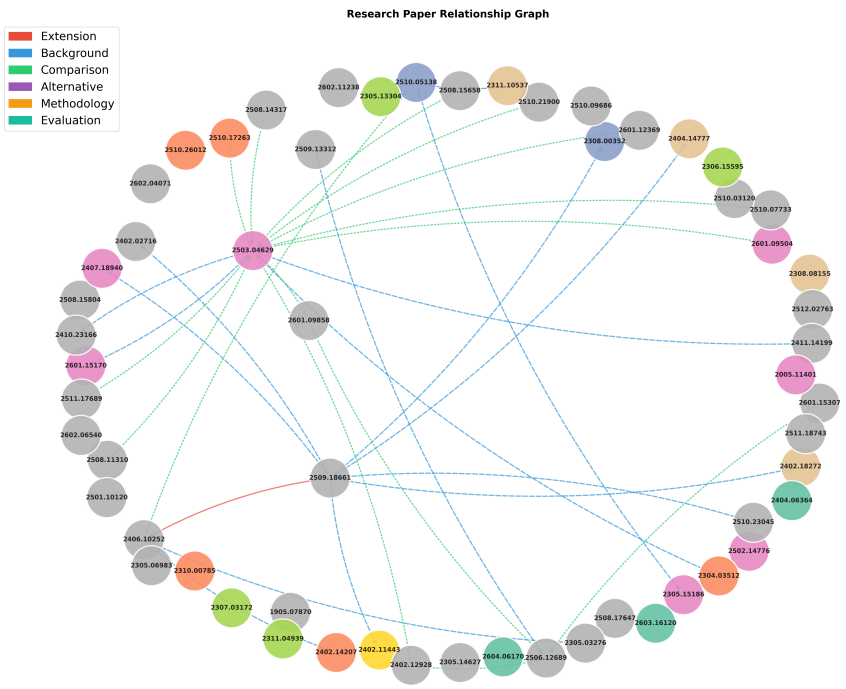


Figure 11: Example of a relation graph among papers

in terms of their ability to capture research evolution and ensure logical flow, and what are the trade-offs in complexity versus performance?

Related Papers:
2601.09504 2510.07733 (2) 2510.21900 (3) 2510.05138 (4) 2509.18661 (5)

Answer: Explicit hierarchical structures (e.g., knowledge trees, citation graphs) directly model research evolution and logical flow. MVSS uses Hierarchical Knowledge Trees to explicitly map conceptual evolution and paradigm shifts, yielding superior structural quality and citation precision <MVSS>. Similarly, SurveyG's layered citation graph (Foundation/Development/Frontier) captures temporal progression, enhancing synthesis and critical analysis <SurveyG>. These approaches ensure strong logical flow but increase complexity and cost (e.g., MVSS costs ∼$0.94/topic).

Iterative refinement methods (e.g., IterSurvey, LiRA) rely on incremental outline updates without explicit models. IterSurvey mimics human research via recurrent outline generation and paper cards, improving coverage and structure iteratively <IterSurvey>. LiRA uses multi-agent workflows for parallel drafting and refinement, focusing on readability and citation accuracy <LiRA>. These methods are more flexible and efficient but may lack explicit evolutionary modeling.

Trade-offs: Explicit hierarchies offer better structural coherence and evolution capture at higher computational cost. Iterative methods are more adaptable and efficient but may produce less explicitly organized narratives.

F.4 Relation Graph Example

We use "AutoSurvey" as the input topic for retrieval, clustering, and graph construction. An example is shown in Figure 11, which illustrates the relationships between different papers under this topic.

F.5 CODE REPORT EXAMPLE

We use "AutoSurvey" as the input topic to generate a code report example, which analyzes data structures, design architectures, and other implementation-specific details. Due to space constraints, only the executive summary and problem-modeling parts are shown.

Code Report Case

```
# Integrated Comprehensive Report: AutoSurvey and Related Papers

## 1. Executive Summary

The synthesis of multiple batch analyses reveals a clear
    evolutionary trajectory in automated survey generation: from
    monolithic, single-pass systems toward **modular, multi-agent,
    iterative frameworks** that mimic human research workflows. The
    dominant implementation pattern is a **hierarchical pipeline**
    combining retrieval-augmented generation (RAG), dynamic
    planning, specialized agent collaboration, and iterative
    refinement. Key architectural elements include:

- **Multi-agent orchestration**: Specialized agents (Planner,
    Writer, Reflector, Verifier) with defined roles and
    communication protocols.
- **Dynamic planning-evidence loops**: Iterative cycles where
    outline generation and evidence acquisition inform each other.
- **Memory-augmented generation**: Persistent memory banks (e.g.,
    paper cards, evidence memory) for targeted retrieval and
    context management.
- **Structured intermediate representations**: Hierarchical
    outlines, citation graphs, and standardized operating
    procedures (SOPs) that guide generation.

**Major gaps** persist in deep understanding, evaluation alignment,
    long-context handling, and domain generalization. The most
    promising future directions involve **neuro-symbolic
    integration**, **adaptive context management**,
    **evaluation-aware generation**, and **collaborative human-AI
    systems** that address these limitations.

## 2. Problem Modeling and Data Structure Classification

A unified view reveals that all systems model survey generation as
    a **hierarchical, evidence-grounded synthesis problem**. The
    core challenge is decomposing a broad topic into a coherent
    structure while integrating and synthesizing information from
    numerous sources.

### Data Structure Categories and Trade-offs

| Category | Specific Examples | Why Chosen | Key Trade-offs |
|--|--|--|--|
| **Hierarchical Outline Structures** | `DynamicOutline`
    (WebWeaver), `HierarchicalDraft` (SciSage), Tree-structured
    outlines (AutoSurvey, AutoSurvey2) | Mirrors human survey
    organization; enables modular, parallel generation and targeted
    retrieval. | **Advantage**: Parallel section generation
    possible; clear quality control per section. **Disadvantage**:
    Rigid structure may not fit all topics; requires accurate
    initial outline; may lose global coherence. |
| **Memory-Augmented Representations** | Natural language "Long
    Short-Term Memory" buffers (RecurrentGPT), "Paper cards"
    distilling contributions (IterSurvey), `EvidenceMemoryBank`
```

```
    (WebWeaver) | Enables long-context generation; provides
    faithful paper-level grounding; reduces redundant processing. |
    **Advantage**: Interpretable, editable memory; targeted
    retrieval. **Disadvantage**: Memory management overhead;
    potential information loss in distillation; retrieval latency. |
| **Agent Communication Structures** | `SOPPrompts` (MetaGPT),
    `AgentMessageQueue` (implied), Role-based prompt engineering |
    Enables structured collaboration between specialized agents;
    reduces hallucination through validation. | **Advantage**:
    Mimics human workflows; clear responsibility boundaries.
    **Disadvantage**: Increased complexity; potential communication
    bottlenecks; may be too rigid for creative tasks. |
| **Retrieval-Augmented & Citation Structures** | Retrieved paper
    corpora with metadata, `CitationGraph` (ReportBench),
    citation-indexed evidence storage | Addresses parametric
    knowledge limitations of LLMs; enables real-time updates and
    citation verification. | **Advantage**: Real-time updates;
    reduces redundancy; improves citation accuracy.
    **Disadvantage**: Retrieval noise; computational cost; memory
    overhead. |
| **Evaluation Structures** | `EvaluationCorpus` (ReportBench),
    `QuestionAnswerPairs` (LooGLE) | Enables systematic
    benchmarking and quality assessment. | **Advantage**:
    Standardized evaluation; reproducible results.
    **Disadvantage**: May not capture all quality dimensions;
    static benchmarks. |
...
```

G Case Study

G.1 Comparison between DeepSurvey and the baseline

To further evaluate the quality and depth of the generated surveys, we compare DeepSurvey with the most influential baseline, AutoSurvey, on the topic "AutoSurvey" (specific examples in Appendix G.3.1 and Figure 12).

Topic focus. DeepSurvey's eight sections remain tightly focused on automated survey generation, with every subsection directly addressing the target topic. In contrast, AutoSurvey's outline includes substantial irrelevant content—Mobile and Sensor-Based Data Collection, Crowdsourcing Platforms, Conversational Survey Design, and domain applications such as Environmental Monitoring and Healthcare—none of which bear meaningful relevance to the topic.

Analytical depth and synthesis. DeepSurvey establishes a four-paradigm taxonomy (one-shot, iterative, multi-agent, structure-first) and identifies a central challenge—the *skeleton-versus-flesh trade-off*—with paper-level references enabling cross-paper comparison along concrete dimensions (e.g., retrieval strategies, refinement loops). AutoSurvey produces a flat enumeration without any taxonomy, unifying framework, or comparative synthesis.

Critical analysis and future directions. DeepSurvey explicitly identifies limitations of existing systems—shallow summaries, citation hallucination, limited domain generalization—and proposes concrete future directions such as hybrid human-AI collaboration and multimodal synthesis. AutoSurvey offers no critical evaluation, no identification of open problems, and no forward-looking analysis.

Structural quality. DeepSurvey provides informative descriptions for each subsection specifying the content scope and analytical angle (e.g., "Reviews foundational approaches that used predefined templates, sentence extraction, and optimization frameworks"). AutoSurvey's subsection titles are

often generic (e.g., "Scalability and Efficiency") and lack scoped descriptions, resembling a textbook table of contents rather than an analytical survey.

G.2 Effect of code-agent subsystem

Beyond the quantitative improvements in content depth (7.66→7.90 in all domains and 7.87→8.28 in CS) (Section 3.5), the code-analysis subsystem recovers implementation-level details that are absent from paper text. Using the topic "Hallucination in LLM" as an example (Case G.3.2), the generated survey integrates code-level findings in three ways: (1) it extracts concrete implementation patterns from repositories—such as spectral-collapse detection via Laplacian eigenvalue monitoring and attention-locking thresholds (∼0.69 JS-Drift saturation)—that are described only abstractly or omitted entirely in the corresponding papers; (2) it embeds implementation sketches (e.g., incremental partial-similarity scoring at each decoding step) that operationalize conceptual mitigation strategies into actionable code; and (3) it synthesizes a unified pipeline class that orchestrates detection, alignment, and evaluation components drawn from multiple papers, bridging the gap between fragmented conceptual discussions and a coherent implementation framework. These code-grounded additions complement the text-based analysis by providing the technological granularity that paper prose alone cannot convey.

G.3 Specific Case

G.3.1 DeepSurvey Survey Case

DeepSurvey's output survey (due to space constraints, we present the abstract and outline rather than the full text) under "AutoSurvey" is as follows:

Generated Survey Case

```
# Automated Literature Survey and Review Generation: A
    Comprehensive Survey

**Abstract**
The exponential growth of scientific literature has created an
    urgent need for automated tools to synthesize knowledge. This
    survey provides a comprehensive analysis of automated survey
    generation systems, tracing their evolution from early
    template-based methods to sophisticated agentic frameworks. We
    identify a central challenge: the 'skeleton-versus-flesh'
    trade-off, where systems excel at organizing outlines and
    citations (the skeleton) but struggle with deep synthesis and
    critical analysis (the flesh). Our survey is structured around
    three pillars: (1) a detailed taxonomy of system architectures,
    including one-shot, iterative, multi-agent, and structure-first
    paradigms; (2) a critical examination of evaluation frameworks,
    highlighting the persistent gap between automated metrics and
    human judgment; and (3) a forward-looking analysis of future
    directions, such as hybrid human-AI collaboration and
    autonomous research pipelines. By synthesizing over fifteen key
    systems and benchmarks, we aim to provide a comparative
    analysis of design philosophies, implementation trade-offs, and
    unresolved challenges, offering a roadmap for advancing
    automated scholarly synthesis.

...

Outline:

## 1. Introduction
Establishes the motivation for automated survey generation by
    highlighting the exponential growth of scientific literature
    and the limitations of manual synthesis. Introduces the core
```

```
    challenges: information overload, context window constraints,
    and the need for structural organization. Positions the
    survey's scope to cover system architectures, evaluation
    paradigms, and future directions, agentic, structure-aware
    generation.          - papers:
    74fdf8053638eb12e5cce073ab4fb476fdb9f9cb

### 1.1 The Information Overload Crisis in Scientific Research
Discusses the exponential growth of publications across
    disciplines, making manual literature reviews impractical and
    outdated quickly. Cites the challenge of synthesizing knowledge
    from hundreds or thousands of papers, leading to redundancy and
    fragmentation in the literature.

### 1.2 From Manual to Automated: The Promise of LLMs
- Introduces Large Language Models (LLMs) as a transformative
    technology for automating survey generation. Highlights their
    ability to process and synthesize text at scale, but notes key
    limitations: context window constraints, parametric knowledge
    cutoffs, and lack of structural reasoning.
- papers: 74fdf8053638eb12e5cce073ab4fb476fdb9f9cb

### 1.3 Core Challenges in Automated Survey Generation
- Outlines the fundamental challenges: (1) Retrieving relevant and
    high-quality papers, (2) Organizing them into a coherent
    hierarchical structure, (3) Synthesizing content with critical
    analysis and proper citations, and (4) Evaluating the quality
    of generated surveys against human standards.

### 1.4 Scope and Organization of This Survey
- Defines the survey's scope: covering system architectures,
    evaluation frameworks, and future directions. Provides a
    roadmap of the subsequent sections, emphasizing the logical
    flow from historical evolution to current systems, evaluation,
    and open problems.

## 2. Historical Evolution and Paradigm Shifts
- Traces the development of automated survey generation from early
    template-based and extractive methods to modern LLM-driven
    systems. Highlights key paradigm shifts: from one-shot
    generation to iterative refinement, from linear pipelines to
    multi-agent architectures, and from text-only outputs to
    multi-view structured representations.
- papers: 2305.13304, 2601.09504, 2510.17263, 2509.18661,
    2602.04071, 2503.04629, 2304.03512, 2510.21900, 2510.26012,
    2502.14776, 2406.10252, 2601.09858, 2508.14317, 2511.17689,
    2506.12689, 99e5d27e6af6f6a53d23ceb0c53e09503c4b6403,
    2510.07733,  213d6264456e3c6410bb9b2e8d7711742bb8eceb

### 2.1 Early Template-Based and Extractive Methods
- Reviews foundational approaches that used predefined templates,
    sentence extraction, and optimization frameworks. Discusses
    their limitations in handling semantic understanding and
    generating novel syntheses, setting the stage for LLM-based
    methods.
- papers: 2304.03512, 99e5d27e6af6f6a53d23ceb0c53e09503c4b6403,
    213d6264456e3c6410bb9b2e8d7711742bb8eceb

### 2.2 The Rise of LLM-Based One-Shot Generation
- Covers the first wave of LLM applications, exemplified by systems
    like AutoSurvey, which used retrieval-augmented generation
    (RAG) and parallel subsection drafting. Highlights their
```

```
    efficiency but notes issues with structural coherence and
    citation accuracy.
- papers: 2503.04629, 2510.26012, 2502.14776, 2406.10252

### 2.3 The Iterative and Agentic Revolution
- Describes the shift towards iterative workflows (e.g.,
    IterSurvey, SurveyGen-I) and multi-agent frameworks (e.g.,
    Agentic AutoSurvey, LiRA). Emphasizes how these systems mimic
    human research processes through incremental retrieval, outline
    refinement, and specialized agent roles.
- papers: 2305.13304, 2509.18661, 2602.04071, 2510.21900,
    2601.09858, 2508.14317, 2511.17689, 2506.12689

### 2.4 Structure-First and Multi-View Paradigms
- Introduces the latest paradigm focusing on explicit structural
    planning before content generation. Highlights systems like
    MVSS (hierarchical knowledge trees), TAXOALIGN (taxonomy
    generation), and SurveyG (citation graphs) that prioritize
    conceptual organization and cross-view consistency.
- papers: 2601.09504, 2510.17263, 2510.07733

## 3. System Architectures and Methodological Frameworks
- Provides a detailed taxonomy of automated survey generation
    systems based on their architectural design. Categorizes
    systems into one-shot, iterative, multi-agent, and
    structure-first paradigms, analyzing their core workflows,
    planning mechanisms, and integration of retrieval and
    generation.
- papers: 2510.05138, 2305.13304, 2601.09504, 2510.17263,
    2509.18661, 2602.04071, 2503.04629, 2511.18743, 2304.03512,
    2510.21900, 2510.26012, 2502.14776, 2406.10252, 2509.13312,
    2601.09858, 2308.00352, 2604.06170, 2508.14317, 2511.17689,
    2506.12689, 99e5d27e6af6f6a53d23ceb0c53e09503c4b6403, 2510.07733

### 3.1 One-Shot Generation Pipelines
- Analyzes systems that generate surveys in a single pass after
    initial retrieval and outline creation. Discusses their typical
    stages: retrieval, outline generation, parallel subsection
    drafting, and integration. Highlights their speed but notes
    limitations in handling long-range dependencies and iterative
    refinement.
- papers: 2503.04629, 2304.03512, 2510.26012, 2502.14776,
    2406.10252, 99e5d27e6af6f6a53d23ceb0c53e09503c4b6403

### 3.2 Iterative and Recurrent Planning Frameworks
- Examines systems that use iterative loops to refine outlines and
    content. Covers mechanisms like recurrent outline generation
    (IterSurvey), dynamic planning with memory (SurveyGen-I), and
    reinforcement learning over outlines (OutlineForge). Emphasizes
    their ability to adapt to discovered knowledge and improve
    coherence over multiple passes.
- papers: 2305.13304, 2511.18743, 2510.21900, 2509.13312,
    2601.09858, 2508.14317

### 3.3 Multi-Agent Collaborative Architectures
- Explores frameworks that decompose the survey task among
    specialized agents (e.g., search, writing, evaluation).
    Discusses coordination mechanisms like publish-subscribe
    (MetaGPT), iterative reflection (SciSage), and rubric-guided
    refinement (ARISE). Highlights their modularity and potential
    for error correction through agent collaboration.
- papers: 2510.05138, 2509.18661, 2602.04071, 2308.00352,
    2604.06170, 2511.17689, 2506.12689
```

```
### 3.4 Structure-First and Multi-View Synthesis
- Focuses on systems that explicitly model hierarchical structures
    before generating text. Covers hierarchical knowledge trees
    (MVSS), taxonomies (TAXOALIGN), citation graphs (SurveyG), and
    attribute trees (SurveyX). Discusses how these structures guide
    outline construction, ensure logical flow, and enable
    cross-view alignment.
- papers: 2601.09504, 2510.17263, 2510.07733

## 4. Core Components and Techniques
- Dissects the key technical components that enable automated
    survey generation: literature retrieval, outline planning,
    content synthesis, and post-processing. Analyzes how different
    systems implement these components, highlighting innovations in
    retrieval strategies, memory mechanisms, and refinement loops.
- papers: 2510.05138, 2305.13304, 2601.09504, 2510.17263,
    2509.18661, 2602.04071, 2503.04629, 2511.18743, 2304.03512,
    2510.21900, 2510.26012, 2502.14776, 2406.10252, 2509.13312,
    2601.09858, 2308.00352, 2508.17647, 2604.06170, 2508.14317,
    2511.17689, 2506.12689,
    74fdf8053638eb12e5cce073ab4fb476fdb9f9cb, 2510.07733

### 4.1 Literature Retrieval and Knowledge Integration
- Covers retrieval strategies from simple keyword search to
    advanced multi-source, multi-granularity approaches. Discusses
    techniques like keyword expansion (SurveyX), co-citation
    analysis (QUAL-SG), and temporal-aware reranking (SurveyForge).
    Highlights the shift from static retrieval to active, iterative
    retrieval based on generation confidence (FLARE).
- papers: 2509.18661, 2503.04629, 2511.18743, 2510.21900,
    2510.26012, 2502.14776, 2406.10252, 2509.13312, 2508.17647,
    2604.06170, 2508.14317, 74fdf8053638eb12e5cce073ab4fb476fdb9f9cb

### 4.2 Outline Planning and Hierarchical Organization
- Analyzes methods for generating survey outlines, from heuristic
    learning from human examples (SurveyForge) to explicit
    hierarchical modeling (MVSS). Discusses the role of knowledge
    trees, taxonomies, and citation graphs in capturing research
    evolution and ensuring logical flow. Compares static vs.
    dynamic outline generation.
- papers: 2305.13304, 2601.09504, 2510.17263, 2503.04629,
    2304.03512, 2510.21900, 2510.26012, 2502.14776, 2509.13312,
    2601.09858, 2508.14317, 2510.07733

### 4.3 Content Synthesis and Citation Grounding
- Examines techniques for generating section content, including
    parallel drafting, memory-driven generation, and hint-based
    synthesis. Focuses on ensuring factual accuracy and proper
    citation through mechanisms like paper cards (IterSurvey),
    evidence audit (RhinoInsight), and RAG-based rewriting
    (SurveyX).
- papers: 2510.05138, 2509.18661, 2503.04629, 2510.21900,
    2510.26012, 2502.14776, 2406.10252, 2509.13312, 2508.14317

### 4.4 Refinement, Alignment, and Post-Processing
- Covers iterative refinement loops that polish surveys for
    coherence, style, and factual consistency. Discusses cross-view
    alignment (MVSS), reviewer-refiner loops (IterSurvey), and
    rubric-guided revision (ARISE). Also addresses post-processing
    steps like figure/table generation and citation verification.
- papers: 2510.05138, 2601.09504, 2602.04071, 2511.18743,
    2510.21900, 2502.14776, 2308.00352, 2511.17689, 2506.12689
```

```
## 5. Evaluation and Benchmarking
- Reviews the diverse evaluation methodologies and benchmarks
    developed to assess automated survey generation systems.
    Contrasts absolute scoring with pairwise ranking, discusses the
    role of LLM-as-a-judge, and highlights the persistent gap
    between automated metrics and human expert judgment. Emphasizes
    the need for discipline-aware and reader-centric evaluation.
- papers: 2510.05138, 2601.09504, 2510.17263, 2509.18661,
    2503.04629, 2602.11238, 2304.03512, 2510.21900, 2510.26012,
    2502.14776, 2406.10252, 2509.13312, 2508.17647, 2508.15804,
    2508.15658, 2511.17689, 2512.02763, 2506.12689, 2510.03120,
    213d6264456e3c6410bb9b2e8d7711742bb8eceb

### 5.1 Automatic Metrics and LLM-as-a-Judge
- Surveys common automatic metrics: ROUGE, BLEU, BERTScore for text
    similarity; citation precision/recall for reference quality;
    and structural metrics like heading recall. Discusses the use
    of LLM-as-a-judge for multi-dimensional assessment (coverage,
    structure, relevance) and its limitations in capturing deep
    synthesis.
- papers: 2601.09504, 2510.17263, 2304.03512, 2510.26012,
    2502.14776, 2406.10252, 2508.17647, 2508.15658, 2511.17689,
    2512.02763, 213d6264456e3c6410bb9b2e8d7711742bb8eceb

### 5.2 Benchmarks for Survey Generation
- Introduces key benchmarks: SciReviewGen (literature reviews),
    SurveyBench (reader-aligned evaluation), SurGE (comprehensive
    CS surveys), and SurveyLens (discipline-aware evaluation).
    Compares their scope, evaluation dimensions, and insights into
    system strengths and weaknesses.
- papers: 2510.05138, 2509.18661, 2503.04629, 2602.11238,
    2510.21900, 2406.10252, 2509.13312, 2508.17647, 2508.15804,
    2508.15658, 2512.02763, 2506.12689, 2510.03120

### 5.3 Human Evaluation and Pairwise Ranking
- Highlights the importance of human evaluation through expert
    ratings, pairwise comparisons (Survey-Arena), and real-user
    studies (MyScholarQA). Discusses how human judgment captures
    nuanced qualities like critical analysis and narrative
    coherence that automated metrics miss.
- papers: 2601.09504, 2602.11238, 2510.21900, 2510.03120

### 5.4 The Evaluation Gap and Future Directions
- Analyzes the persistent gap between automated metrics and human
    judgment, particularly in assessing synthesis depth and
    academic value. Proposes future directions: hybrid evaluation
    paradigms, discipline-specific rubrics, and quiz-based
    answerability tests to better align with reader needs.
- papers: 2602.11238, 2508.15804, 2512.02763, 2510.03120

## 6. Challenges and Limitations
- Identifies the major unresolved challenges in automated survey
    generation, including the 'skeleton-versus-flesh' trade-off,
    citation hallucination, domain generalization, and the lack of
    deep synthesis. Discusses how current systems fall short of
    human experts in critical analysis, novelty, and scholarly
    communication.
- papers: 2511.18743, 2602.11238, 2508.17647, 2508.15804, 2508.15658

### 6.1 The Skeleton-Versus-Flesh Trade-off
- Explores the fundamental tension between structural proficiency
    and content depth. Notes that while systems excel at organizing
```

```
    outlines and citations, they often produce shallow summaries
    lacking critical analysis, novel insights, and nuanced
    argumentation. Highlights the gap in 'academic value' metrics.

### 6.2 Citation Accuracy and Hallucination
- Discusses persistent issues with citation quality: hallucinated
    references, over-citation, and misalignment between claims and
    sources. Analyzes architectural constraints (evidence locking,
    frozen structures) and verification mechanisms that mitigate
    but do not fully solve the problem.
- papers: 2511.18743, 2508.15804

### 6.3 Domain Generalization and Adaptability
- Examines the limited generalizability of current systems, which
    are often trained and evaluated on computer science topics.
    Highlights challenges in adapting to other disciplines with
    different writing norms, citation practices, and structural
    conventions. Notes the STEM bias in existing benchmarks.
- papers: 2602.11238

### 6.4 Lack of Deep Synthesis and Critical Analysis
- Identifies the core limitation: LLMs struggle to perform genuine
    scholarly synthesis---integrating findings across papers,
    identifying research gaps, and offering novel perspectives.
    Discusses how iterative refinement and multi-agent debate
    improve but do not resolve this limitation.
- papers: 2508.17647, 2508.15658

## 7. Future Directions and Open Problems
- Outlines specific, actionable research directions to advance
    automated survey generation. Emphasizes the need for hybrid
    human-AI collaboration, improved structural reasoning,
    multimodal synthesis, and robust evaluation frameworks.
    Proposes concrete steps for addressing the identified
    challenges.
- papers: 2305.13304, 2602.04071, 2511.18743, 2601.09858,
    2604.06170, 2404.06364

### 7.1 Hybrid Human-AI Collaboration and Interactive Systems
- Advocates for integrating human guidance at critical stages:
    reference selection, outline approval, and iterative
    refinement. Proposes interactive interfaces that allow
    researchers to steer generation, correct errors, and inject
    domain knowledge, moving beyond fully automated pipelines.
- papers: 2305.13304, 2602.04071, 2511.18743, 2604.06170, 2404.06364

### 7.2 Advanced Structural Reasoning and Planning
- Calls for research into explicit structural inductive biases,
    such as reinforcement learning over outlines, symbolic
    reasoning for taxonomy construction, and dynamic knowledge
    graphs. Emphasizes the need for models that can infer and
    represent expert knowledge hierarchies from literature.
- papers: 2601.09858

### 7.3 Multimodal Synthesis and Rich Media Integration
- Highlights the gap in generating non-textual elements: figures,
    tables, equations, and interactive visualizations. Proposes
    developing systems that can automatically create and integrate
    these elements to enhance survey readability and information
    density.

### 7.4 Robust, Discipline-Aware Evaluation Frameworks
```

```
- Advocates for evaluation paradigms that combine automated metrics
    with human-aligned benchmarks, discipline-specific rubrics, and
    reader-centric tests (e.g., quiz-based answerability).
    Emphasizes the need for metrics that assess academic value,
    critical analysis, and personalization.

### 7.5 Towards Autonomous Research Pipelines
- Envisions the integration of survey generation into broader AI
    Scientist systems that can conduct end-to-end research: from
    literature review to idea generation, experimentation, and
    paper writing. Identifies missing capabilities: experimental
    execution, uncertainty quantification, and cross-domain
    transfer.
- papers: 2602.04071

## 8. Conclusion
- Summarizes the survey's key findings: the rapid evolution from
    simple pipelines to sophisticated agentic systems, the
    persistent challenges in deep synthesis and evaluation, and the
    promising future directions. Reiterates the transformative
    potential of automated survey generation while acknowledging
    the need for continued human oversight and interdisciplinary
    collaboration.
```

G.3.2 DEEPSURVEY WITH CODE-AGENT SUBSYSTEM CASE

DeepSurvey's output survey under "Hallucination in LLM" is as follows:

DeepSurvey case: code-agent subsystem's improvements

```
...

To deploy this surgically in production, the Code Report (Section
    7, Direction 2) suggests an incremental scoring approach that
    operates at each decoding step. This method computes a partial
    similarity score between generated tokens and a reference,
    triggering activation cancellation if confidence drops below a
    threshold. The real-time hook, as implemented in [8], can be
    optimized with computational techniques from the Code Report
    (Section 4), such as caching intermediate hidden states and
    batch processing multiple candidate tokens, reducing latency
    while preserving the surgical precision of H-Node suppression.

**Implementation Sketch:**

# At each decoding step:
partial_similarity = compute_partial_similarity(
    generated_tokens[:current_step], reference_tokens
)
if partial_similarity < confidence_threshold:
    suppress_h_node_activations()

...

To address the fragmentation noted in the Code Report---where
    evaluation, detection, and alignment operate in isolation---a
    unified pipeline integrating [10] for evaluation, [34] for
    detection, and [26] for alignment could provide a holistic
    framework. For instance, a class-based implementation could
    orchestrate these components: during generation, real-time
    uncertainty from [34] triggers alignment corrections via [26],
```

```
    followed by [10] evaluation of the refined output, as sketched
    in Direction 1 of the Code Report. This directly counters the
    'dimensional poverty' critique by demonstrating how
    multi-dimensional scoring can be operationalized in a single,
    actionable system. Concretely, [9]'s six-stage pipeline (task
    formulation, decomposition, tool-augmented execution,
    structured generation, multi-dimensional scoring, trace
    analysis) could be extended with a concrete class
    implementation inspired by the Code Report's Direction 1:


class UnifiedHallucinationPipeline:
def __init__(self, evaluator=BERTScore(), detector=SpikeScore(),
    aligner=VaryingShadesOfWrong()):
    self.evaluator = evaluator
    self.detector = detector
    self.aligner = aligner
def process(self, query, context):
    # (1) Task Formulation: Parse user query and context
    # (2) Decomposition: Break query into sub-tasks using LENS's
    tree-based query decomposition
    # (3) Tool-Augmented Execution: Generate responses while
    SpikeScore monitors uncertainty fluctuations across turns
    responses = self.detector.track_uncertainty(query, context)
    if self.detector.hallucination_flag:
        # (4) Structured Generation: Trigger alignment corrections
    via preference optimizatio
        corrected = self.aligner.correct(responses, query)
    else:
        corrected = responses
    # (5) Multi-Dimensional Scoring: Compute BERTScore F1,
    faithfulness, coverage` `        score =
    self.evaluator.compute(corrected, context)
    # (6) Trace Analysis: Log causal attribution via OpenTelemetry
    return score, corrected


    This pipeline operates as follows: For a medical diagnosis
    query, decomposition yields symptom extraction, evidence
    retrieval, and synthesis; [34] detects hallucination spikes
    during synthesis; [26] corrects using wrong-over-wrong
    preferences; [10] evaluates final output against retrieved
    evidence. This integration not only operationalizes [9]'s
    framework but also addresses Code Report gaps by enabling
    real-time detection and alignment, as validated in case studies
    where such pipelines reduced hallucination rates by 30-40\% in
    high-stakes domains.

...
```

H PROMPT ENGINEER

In this section, we present the core prompts used in the DeepSurvey agentic system, including outline generation and drafting. The remaining prompts can be found in our anonymous github's link or submitted supplementory material.

H.1 OUTLINE PROMPT

This part demonstrates the outline creation/refinement prompt.

Outline Creation Prompt

```
SURVEY_OUTLINE_GENERATION_OUTLINE_DRAFT = """You are an expert
    research survey generator. Your task is to generate and
    iteratively update an existing survey outline using a batch of
    new paper keynotes, the current outline, and the analysis
    results of relevant papers for the topic/subtopic being written.

**Guidance:**
- The analysis results contain summarized insights, comparisons,
    trends, and key points extracted from prior work. They should
    be the **important source of guidance** when updating the
    outline.
- Paper keynotes are also useful: use them to **supplement,
    validate, or provide additional details** for the
    sections/subsections.
- Balance both sources to create a comprehensive, accurate, and
    logically organized survey outline.

**Requirements:**
1. The outline should contain multiple sections, subsections and
    their descriptions.
2. The output outline should have excellent organization and meet
    academic standards.
3. The outline should exhibit excellent rigor: ensuring that the
    content of each subsection falls within the scope of the
    current section.
4. The outline should ensure that it covers a wide range of content
    under the topic while staying within the scope of the topic.
5. Use the current outline as the base structure. Keep existing
    sections/subsections unless updated or merged. If the current
    outline is empty, create a new outline from scratch.
6. The outline should contain a **Conclusion** section and a
    **Future Work** section/subsection.
7. The outline should exhibit good logic to ensure the entire
    survey flows smoothly.
8. Ensure a balanced number of subsections in the main sections
    (excluding the conclusion and introduction).
9. Make sure most of the corresponding new paper in **new paper
    keynotes** can be included in at least one subsection or
    section of the outline.
10. You are provided with other relevant papers which are retrieved
    from the database, you can use them to better understand and
    generate.
11. Maintain clarity, logical structure, and a survey-style
    narrative.
12. Ensure logical coherence between the sections, avoiding
    excessive independence and fragmentation. For instance, do not
    add a "Conclusion" subsection to every section, which leads to
    logical fragmentation between different sections.
13. Output strictly in JSON format, as shown below.

**Input:**
- current outline: {current_outline}

- key papers: {paper_keynotes}

- key papers analysis: {papers_analysis}

- other relevant papers: {other_relevant_papers}

**Output JSON format:**
{{
```

```
    "title" : "Survey_Title",
    "sections": [
        {{
            "title": "Section_title",
            "description": "Summary of content to include,
    emphasizing insights, comparisons, and trends from analysis,
    supplemented by keynotes",
            "subsections": [
                {{
                    "title": "Subsection_title",
                    "description": "Content description reflecting
    key insights, trends, comparisons from analysis, supplemented
    by keynotes",
                }}
            ]
        }}
]
}}

**Instruction to LLM:**
- Use the insights, trends, and comparisons from relevant paper
    analysis together with points from new paper keynotes to update
    the outline. Both sources should inform the content of each
    section/subsection.
- Add, merge, or revise sections/subsections as appropriate,
    keeping the outline structured and coherent.
- Ensure that each section/subsection reflects the combined
    contributions of the analysis results and the keynotes,
    capturing important insights, trends, and examples.
- Check whether the input outline satisfies the requirements and
    update to meet the requirements if not.
- Only generate the outline in the required JSON format. Do not
    include specific paper IDs in the outline.
"""
```

H.2 DRAFT PROMPT

In this part we will present the prompt for section draft and subsection draft.

Subsection Draft Prompt

```
SUBSECTION_DRAFT = """You are an expert in writing surveys that
    meet the standards of top academic conferences. The ultimate
    goal is to complete an academic survey with depth and insights
    that can boost further development, rather than simply listing
    methods.

Now you are responsible for writing a subsection of the survey
    paper.

**Subsection Title**:
{title}

**Guidance**:
Write a coherent subsection that explains, analyzes, or discusses
    the topic indicated in the title and description.
You should synthesize relevant information from the papers,
    analysis results.
The content should be academically structured and readable, with
    emphasis on insights, trends, and comparisons where appropriate.
```

```
You may include examples from papers to support the discussion, but
    do not simply list papers.
You are encouraged to cite more papers from the relevant papers to
    strengthen your points.
You are provided with the outline of the whole survey. Make sure
    the subsection content coherent to the survey logic.
Utilize the input content in a safe and reasonable manner, and
    ensure that the readability, structure, and depth of the
    content of the paragraph meet the requirements of a top-tier
    conference survey.

{code_report_prompt}

**Information to use**:
- Subsection description:
{description}

- Closely Relevant papers:
{papers}

- Other Relevant papers:
{other_relevant_papers}

- Survey Outline:
{survey_outline}

- Insights from analysis:
{relevant_analysis}

**Output:**
- A well-written subsection in several coherent paragraphs.
- Academic style; focused on synthesis and analysis, not just
    reporting.
- **Only** cite papers that appear in the provided input.
- Strictly use format: <paper_title> (like <Attention is All You
    Need>) to cite wherever appropriate.
- Cite at least **{subsection_least_citations} different papers**
    for a subsection. You are encouraged to cite more papers to
    give in-depth analysis.
- Do not generate a bibliography or reference list here.
- The content of each subsection should be at least
    {subsection_least_words} words long
- Generate the content directly. DO NOT generate any subsection
    title or section header here.
- CRITICAL: '#' is used for section/subsection anchor. Avoid any
    '#' in the output content.
"""
```

Section Draft Prompt

```
SECTION_DRAFT = """You are writing the Introductory
    Paragraphs/Preamble for a specific section of an academic
    top-tier conference survey paper.
**Task**:
Write the opening text that appears immediately after the Section
    Title but before the first subsection. This content should
    serve as a high-level synthesis and roadmap for the reader.

**Requirements**:
1. Synthesize, Don't Summarize: Do not simply list what each
    subsection will do. Instead, explain the logic behind why this
```

```
    section is structured this way and the significance of these
    topics within the broader field.
2. Establish Definition & Scope: Clearly define the core concepts
    covered in this section. Refer to <paper_title> to establish
    foundational definitions or taxonomies.
3. Identify Trends: Highlight the overarching trends or challenges
    that link the following subsections together.
4. Academic Tone: Maintain a formal, authoritative, and objective
    voice.
5. Citations: * Only cite papers provided in the relevant papers
    provided below(Closely Relevant papers and Other Relevant
    papers).
Use the format: <paper_title>.
6. Cite at least {section_least_citations} papers to ground the
    section's scope.
7. Utilize the input content in a safe and reasonable manner, and
    ensure that the readability, structure, and depth of the
    content of the paragraph meet the requirements of a top-tier
    conference survey.
8. Length: The introductory content must be at least
    {section_least_words} words long but avoid too-long preamble
    which will destroy structure and readability.
9. CRITICAL: '#' is used for section/subsection anchor. Avoid any
    '#' in the output content.

**Input**:
- Section Title:
{title}

- Section Description:
{description}

- Full Survey Outline:
{survey_outline}

Planned Subsections under this Section: {subsection_drafts} (Use
    these as a roadmap, do not detail their findings yet).

- Closely Relevant papers:
{papers}

- Other Relevant papers:
{other_relevant_papers}

**Output**:
Generate the introductory content directly. DO NOT generate any
    subsection title or section header here.
"""
```

Table 19: Citation retrieval performance. Each value is reported as mean ± standard deviation (coefficient of variation).

Method	Precision	Recall
DeepSurvey	**0.569 ± 0.050 (8.7%)**	**0.630 ± 0.047 (7.5%)**
AutoSurvey	0.342 ± 0.004 (1.2%)	0.440 ± 0.002 (0.5%)
SurveyX	0.125 ± 0.006 (4.8%)	0.194 ± 0.014 (7.4%)
SurveyForge	0.421 ± 0.029 (6.8%)	0.503 ± 0.050 (9.9%)
LiRA	0.227 ± 0.010 (4.3%)	0.442 ± 0.009 (2.1%)

Table 20: Per-metric stability statistics across models, methods, and evaluation dimensions.

Model	Setting	Metric	Std	Range	CV%	Max Abs Dev
gpt-exp	AutoSurvey	Content_Depth	0.064	0.125	0.748	0.069
gpt-exp	DeepSurvey	Content_Depth	0.042	0.083	0.481	0.042
gpt-no-exp	AutoSurvey	Content_Depth	0.032	0.062	0.379	0.035
gpt-no-exp	DeepSurvey	Content_Depth	0.043	0.083	0.511	0.049
mimo-exp	AutoSurvey	Content_Depth	0.012	0.021	0.152	0.014
mimo-exp	DeepSurvey	Content_Depth	0.000	0.000	0.000	0.000
mimo-no-exp	AutoSurvey	Content_Depth	0.043	0.083	0.537	0.049
mimo-no-exp	DeepSurvey	Content_Depth	0.075	0.146	0.927	0.083
gpt-exp	AutoSurvey	Core_Quality	0.087	0.167	1.014	0.097
gpt-exp	DeepSurvey	Core_Quality	0.064	0.125	0.765	0.069
gpt-no-exp	AutoSurvey	Core_Quality	0.024	0.042	0.289	0.028
gpt-no-exp	DeepSurvey	Core_Quality	0.064	0.125	0.781	0.069
mimo-exp	AutoSurvey	Core_Quality	0.024	0.042	0.285	0.028
mimo-exp	DeepSurvey	Core_Quality	0.208	0.417	2.994	0.208
mimo-no-exp	AutoSurvey	Core_Quality	0.072	0.125	0.875	0.083
mimo-no-exp	DeepSurvey	Core_Quality	0.024	0.042	0.306	0.028
gpt-exp	AutoSurvey	Total_Score	0.045	0.087	0.527	0.049
gpt-exp	DeepSurvey	Total_Score	0.038	0.075	0.452	0.039
gpt-no-exp	AutoSurvey	Total_Score	0.015	0.029	0.176	0.015
gpt-no-exp	DeepSurvey	Total_Score	0.031	0.062	0.382	0.033
mimo-exp	AutoSurvey	Total_Score	0.020	0.038	0.239	0.022
mimo-exp	DeepSurvey	Total_Score	0.133	0.267	1.822	0.135
mimo-no-exp	AutoSurvey	Total_Score	0.029	0.050	0.355	0.033
mimo-no-exp	DeepSurvey	Total_Score	0.010	0.021	0.133	0.011
gpt-exp	AutoSurvey	Writing_Quality	0.032	0.062	0.394	0.035
gpt-exp	DeepSurvey	Writing_Quality	0.024	0.042	0.302	0.028
gpt-no-exp	AutoSurvey	Writing_Quality	0.021	0.042	0.255	0.021
gpt-no-exp	DeepSurvey	Writing_Quality	0.012	0.021	0.147	0.014
mimo-exp	AutoSurvey	Writing_Quality	0.042	0.083	0.515	0.042
mimo-exp	DeepSurvey	Writing_Quality	0.043	0.083	0.573	0.049
mimo-no-exp	AutoSurvey	Writing_Quality	0.032	0.062	0.404	0.035
mimo-no-exp	DeepSurvey	Writing_Quality	0.043	0.083	0.565	0.049

Table 21: Fine-grained ablation scores across Core Quality, Writing Quality, and Content Depth. "w/o X" denotes removal of module X (coll.=collector, analy.=analyzer, keyn.=keynote, writ.=writer, ref.=refinement, agen.=agentic); "DS" denotes DeepSurvey.

Metric	w/o coll.	w/o analy.	w/o keyn.	w/o writ.	w/o ref.	w/o agen.	w/o code	DS
Core Quality								
Synthesis Quality	8.333	7.667	8.222	7.667	8.323	8.209	8.333	**8.334**
Organization	7.329	8.533	8.667	7.333	8.123	8.900	8.833	**8.935**
Comprehensiveness	3.333	8.857	8.714	8.000	8.667	8.652	8.667	**8.889**
Relevance	1.333	9.000	8.933	8.933	8.877	9.247	**9.500**	9.300
Writing Quality								
Readability	7.333	7.333	7.563	7.433	7.431	7.571	**7.830**	7.711
Academic Rigor	6.000	8.333	8.345	5.333	8.153	8.143	8.000	**8.506**
Clarity & Coherence	4.333	8.162	8.333	8.333	7.528	8.333	**8.500**	8.333
Content Depth								
Critical Analysis	6.967	7.121	7.200	5.341	7.162	7.023	7.286	**7.389**
Novelty and Insights	7.633	7.111	6.667	6.000	6.948	7.635	7.533	**7.866**
Specificity	6.942	7.286	7.671	7.495	8.003	7.914	7.833	**8.194**
Future Directions	6.667	7.867	7.333	6.699	7.333	7.915	7.967	**8.139**

Table 22: Ablation results of multi-granularity refinement on Core Quality, Writing Quality, and Content Depth. "w/o Sur./Sub./Sec." denotes removal of survey-, subsection-, and section-level refinement, respectively; "w/o All" removes all three.

Metric	Full	w/o Sec.	w/o Sub.	w/o Sur.	w/o All
Core Quality					
Synthesis Quality	9.067	9.000	9.000	9.000	9.000
Organization	8.867	9.000	9.000	9.000	9.000
Comprehensiveness	9.000	8.833	9.000	9.000	9.000
Relevance	9.400	9.333	9.000	9.200	9.429
Writing Quality					
Readability	8.000	8.000	8.000	8.000	8.000
Academic Rigor	8.933	9.000	9.000	9.000	7.430
Clarity & Coherence	8.000	8.167	8.000	8.000	8.000
Content Depth					
Critical Analysis	8.933	8.667	8.800	8.600	8.286
Novelty and Insights	8.400	8.167	8.200	8.000	8.143
Specificity	7.933	7.500	7.200	7.400	7.000
Future Directions	8.533	8.167	8.000	8.000	8.000

Table 23: DeepSurvey system parameters.

Module / Stage	Parameter	Value	Description
General	Max context length	512k tokens	Per-call input context limit
General	Generation Model	MiMo-V2.5	Model for survey generation
General	Judge Model	MiMo-V2.5 (primary); MiMo-V2-Flash, GPT-5-mini (supplementary)	Model for evaluation
General	Embedding Model	MiniLMv2	Pretrained embedding model
General	Repeated runs	4	Number of runs averaged for the main experiment
Retrieval	Max seed papers	15	Seed papers retained from initial search
Retrieval	Graph expansion depth	1	Max citation graph traversal layers
Retrieval	Papers per seed	20	Max expanded papers retained per seed
Outline generation	Temperature	0.5	Controls outline diversity
Subsection drafting	Temperature	0.7	Controls subsection writing diversity
Section drafting	Temperature	0.3	Controls section integration stability
Pseudocode gen.	Planner temperature	0	Deterministic planning
Pseudocode gen.	Reviewer temperature	0	Deterministic review
Pseudocode gen.	Reviser temperature	0	Deterministic revision
Pseudocode gen.	Creator temperature	0.3	Moderate generation flexibility
Pseudocode gen.	Max rounds	10	Iteration limit for pseudocode generation
Section refinement	Planner temperature	0.7	Section-level revision planning
Section refinement	Reviewer temperature	1.0	Broad coverage in review
Section refinement	Reviser temperature	0.5	Balanced revision diversity
Subsection refinement	Temperature	0.1	Local stable revision
Survey refinement	Planner temperature	0.7	Survey-level revision planning
Survey refinement	Reviewer temperature	1.0	Global issue identification
Survey refinement	Reviser temperature	0.5	Controllable global revision
Survey refinement	Max rounds	5	Iteration limit for survey refinement



Figure 12: Case Study: DeepSurvey vs. AutoSurvey – DeepSurvey Surpasses AutoSurvey with Closely Topic-Focused Content, Clear Structure, In-Depth Analysis, and Meaningful Future Directions

Table 24: Overall evaluation results. Each value is reported as mean ± standard deviation (coefficient of variation).

Method	Total_Score	Core_Quality	Writing_Quality	Content_Depth
DeepSurvey	**8.341 ± 0.066 (0.8%)**	**8.865 ± 0.129 (1.5%)**	**8.183 ± 0.169 (2.1%)**	**7.897 ± 0.172 (2.2%)**
AutoSurvey	7.958 ± 0.173 (2.2%)	8.646 ± 0.254 (2.9%)	8.153 ± 0.197 (2.4%)	7.171 ± 0.088 (1.2%)
Survey	6.993 ± 0.036 (0.5%)	8.187 ± 0.058 (0.7%)	7.067 ± 0.319 (4.5%)	5.763 ± 0.209 (3.6%)
SurveyForge	8.071 ± 0.106 (1.3%)	8.616 ± 0.046 (0.5%)	8.153 ± 0.188 (2.3%)	7.485 ± 0.146 (1.9%)
LiRA	5.687 ± 0.184 (3.2%)	5.049 ± 0.316 (6.3%)	6.805 ± 0.233 (3.4%)	5.766 ± 0.203 (3.5%)

Table 25: Fine-grained evaluation results for Core Quality.

Method	Synthesis Quality	Organization	Comprehensiveness	Relevance
DeepSurvey	**8.334 ± 0.441 (5.3%)**	**8.935 ± 0.172 (1.9%)**	**8.889 ± 0.255 (2.9%)**	9.300 ± 0.096 (1.0%)
AutoSurvey	7.983 ± 0.440 (5.5%)	8.608 ± 0.254 (3.0%)	8.718 ± 0.347 (4.0%)	9.276 ± 0.096 (1.0%)
SurveyX	7.283 ± 0.417 (5.7%)	8.133 ± 0.503 (6.2%)	8.000 ± 0.200 (2.5%)	**9.333 ± 0.306 (3.3%)**
SurveyForge	7.975 ± 0.248 (3.1%)	8.700 ± 0.188 (2.2%)	8.690 ± 0.207 (2.4%)	9.100 ± 0.111 (1.2%)
LiRA	7.850 ± 0.719 (9.2%)	8.425 ± 0.284 (3.4%)	2.565 ± 0.265 (10.3%)	1.355 ± 0.042 (3.1%)

Table 26: Fine-grained evaluation results for Writing Quality.

Method	Readability	Academic Rigor	Clarity & Coherence
DeepSurvey	7.711 ± 0.188 (2.4%)	**8.506 ± 0.258 (3.0%)**	8.333 ± 0.167 (2.0%)
AutoSurvey	7.606 ± 0.517 (6.8%)	8.500 ± 0.167 (2.0%)	8.353 ± 0.095 (1.1%)
SurveyX	7.053 ± 0.117 (1.7%)	6.313 ± 0.669 (10.6%)	7.833 ± 0.497 (6.3%)
SurveyForge	**7.793 ± 0.190 (2.4%)**	8.283 ± 0.315 (3.8%)	**8.382 ± 0.097 (1.2%)**
LiRA	7.695 ± 0.440 (5.7%)	5.155 ± 0.512 (9.9%)	7.565 ± 0.745 (9.8%)

Table 27: Fine-grained evaluation results for Content Depth.

Method	Critical Analysis	Novelty & Insights	Specificity	Future Directions
DeepSurvey	**7.389 ± 0.347 (4.7%)**	**7.866 ± 0.106 (1.4%)**	**8.194 ± 0.514 (6.3%)**	**8.139 ± 0.097 (1.2%)**
AutoSurvey	7.049 ± 0.599 (8.5%)	6.847 ± 0.342 (5.0%)	7.991 ± 0.256 (3.2%)	6.798 ± 0.654 (9.6%)
SurveyX	5.733 ± 0.286 (5.0%)	5.133 ± 0.222 (4.3%)	5.983 ± 0.563 (9.4%)	6.200 ± 0.388 (6.3%)
SurveyForge	7.350 ± 0.418 (5.7%)	7.290 ± 0.552 (7.6%)	7.895 ± 0.252 (3.2%)	7.405 ± 0.315 (4.3%)
LiRA	5.775 ± 0.587 (10.2%)	5.045 ± 0.349 (6.9%)	6.095 ± 0.564 (9.3%)	6.150 ± 0.412 (6.7%)

Table 29: Detailed scoring rubric for **Core Quality**.

Score	Level Descriptor
Synthesis Quality — distinguishes true integration and conceptual grouping of literature from mere enumeration of papers.	
1	No synthesis — the text is a disconnected list of papers with no comparison, grouping, or integrative statements.
2	Very weak synthesis — occasional grouping but mostly descriptive lists; links between works are absent or mistaken.
3	Minimal synthesis — some attempts to group papers or note similarities, but links are shallow and inconsistent.
4	Basic synthesis — groups and comparisons exist but are superficial and miss deeper patterns or tensions.
5	Moderate synthesis — reasonable grouping and some comparative statements; important cross-paper themes are noted but not developed.
6	Good synthesis — clear groups and comparisons, with emerging themes and some cross-cutting analysis.
7	Strong synthesis — integrates findings across multiple works, highlights causes of differences, and builds partial conceptual links.
8	Very strong synthesis — conveys nuanced relationships among studies, articulates patterns and contradictions, and forms clear conceptual maps.
9	Excellent synthesis — deeply integrates diverse sources, proposes coherent frameworks, and resolves or explains major discrepancies.
10	Outstanding synthesis — creates novel unifying frameworks or taxonomies, transforms raw literature into new conceptual insight.

Continued on next page

Table 29 – *Continued from previous page*

Score	Level Descriptor
Organization — evaluates logical flow, section/subsection ordering, and ease of following the argument across the survey.	
1	No logical order — sections are chaotic and the reader cannot form a coherent view of the topic.
2	Very poor ordering — frequent abrupt jumps and misplaced content; reader must constantly backtrack.
3	Weak organization — some logical chunks but many misplaced paragraphs or unclear section boundaries.
4	Below-average organization — parts make sense but transitions are rough and structure sometimes redundant.
5	Adequate organization — overall structure is serviceable though some sections could be rearranged for clarity.
6	Good organization — clear sections and logical progression with occasional weak transitions or minor redundancies.
7	Strong organization — well-ordered sections, clear flow, and purposeful subsectioning.
8	Very strong organization — sections build on each other smoothly; transitions are purposeful and guide the reader.
9	Excellent organization — highly coherent structure, logical sequencing, and efficient, non-redundant layout.
10	Exemplary organization — impeccable flow, perfect chapter/subsection design, and the structure itself aids understanding.
Comprehensiveness — breadth of coverage across subtopics, methods, datasets, and perspectives relevant to the research area.	
1	Extremely narrow — only a tiny slice of the topic covered; many core areas absent.
2	Very limited — major subtopics missing; coverage skewed to a few papers or approaches.
3	Sparse — several important areas omitted and coverage lacks depth.
4	Partial — some important areas covered but many relevant subtopics absent or lightly treated.
5	Moderate — key areas present but notable gaps in methods, datasets, or perspectives.
6	Reasonably comprehensive — most core areas covered though some peripheral topics missing.
7	Comprehensive — broad coverage including major methods and datasets with minor omissions.
8	Very comprehensive — covers almost all relevant facets with useful detail across areas.
9	Extensive — deep and wide coverage across subfields and perspectives; only small gaps remain.
10	Exhaustive — near-complete coverage across the domain, methods, datasets, and viewpoints.
Relevance — measures how well the content of the survey aligns with the research topic and maintains a clear, unwavering focus throughout. It evaluates whether every section, paragraph, and sentence contributes to advancing the reader's understanding of the stated topic, and whether tangential content is either absent or minimal.	
1	Completely irrelevant — the survey addresses a fundamentally different topic; content has no meaningful connection to the stated research area.
2	Mostly irrelevant — the survey only tangentially touches the topic, with the majority of content being off-topic, peripheral, or unrelated.
3	Weakly relevant — several sections or subsections discuss the topic, but significant portions (more than 30%) are digressive or off-topic.
4	Somewhat relevant — the survey covers the topic but often digresses into unrelated areas; focus is inconsistent throughout.
5	Moderately relevant — the survey is generally on topic, with most content aligning with the research focus, though notable unrelated details appear in places.
6	Reasonably relevant — the survey maintains focus on the topic with occasional minor digressions; most sections contribute meaningfully to the core subject.
7	Strongly relevant — the survey is largely focused, with consistent alignment to the research topic; only rare and minor unrelated content present.
8	Very strongly relevant — the survey is tightly focused with near-perfect topical alignment; every section contributes to the core subject with only negligible peripheral content.
9	Exceptionally relevant — the survey demonstrates exemplary topical coherence; content is exclusively focused on the research topic with no discernible digressions.

Continued on next page

Table 29 – *Continued from previous page*

Score	Level Descriptor
10	Perfectly relevant — the survey is a model of topical unity; every single element, from introduction to conclusion, is laser-focused on advancing the reader's understanding of the research topic, with no off-topic content whatsoever.

Table 30: Detailed scoring rubric for **Writing Quality**.

Score	Level Descriptor
Readability — clarity and accessibility for the intended academic audience (sentence-level clarity, pacing, and jargon handling).	
1	Unreadable — sentences are confusing, grammar errors abound, and jargon is opaque.
2	Very poor readability — frequent grammar issues and unclear sentences; reader struggles to extract meaning.
3	Low readability — many awkward sentences and unclear phrasing; frequent re-reading required.
4	Below-average readability — understandable in places but style is verbose or clumsy; jargon often unexplained.
5	Fair readability — generally clear though some sentences or paragraphs are convoluted.
6	Good readability — clear prose overall with occasional dense passages or uneven pacing.
7	Very good readability — fluent writing, appropriate use of terminology, and easy to follow.
8	Excellent readability — crisp sentences, well-balanced pacing, and accessible to the target academic audience.
9	Near-perfect readability — highly engaging academic prose with precise phrasing and smooth flow.
10	Outstanding readability — exemplary clarity and elegance; highly accessible without loss of technical precision.
Academic Rigor — fidelity to scholarly standards (proper citations, fair representation of prior work, and methodological transparency).	
1	No rigor — citations missing or blatantly incorrect; claims unsupported and methods misrepresented.
2	Very low rigor — many unsupported claims, poor citation practice, and factual errors.
3	Low rigor — partial or inconsistent citation and occasional misrepresentation of methods/results.
4	Below-average rigor — citations present but incomplete and some claims lack supporting evidence.
5	Moderate rigor — generally correct citations and basic methodological descriptions but missing nuance.
6	Good rigor — solid citation practice and fair representation of methods with minor gaps.
7	Strong rigor — careful referencing, transparent presentation of methods and limitations.
8	Very strong rigor — thorough citations, balanced critique of methods, and clear evidence-backed claims.
9	Excellent rigor — meticulous referencing, rigorous methodological critique, and consistently supported claims.
10	Exceptional rigor — exemplary scholarship: exhaustive, accurate citations and deep, transparent methodological analysis.
Clarity & Coherence — the extent to which the survey presents technical ideas with precision, unambiguous definitions, and internally consistent terminology, claims, and conclusions across sections.	
1	Very poor — highly vague, undefined terms, and multiple contradictions or inconsistent claims throughout.
2	Poor — frequent ambiguity and major inconsistencies that seriously hinder understanding.
3	Weak — many passages are unclear, with noticeable drifting terminology or conflicting statements.
4	Below average — some key ideas are understandable, but clarity is limited and inconsistencies remain in several places.
5	Moderate — the survey is mostly understandable, but important technical details are underspecified and coherence is uneven.

Continued on next page

Table 30 – *Continued from previous page*

Score	Level Descriptor
6	Fairly good — generally clear and mostly consistent, with a few imprecise explanations or minor inconsistencies.
7	Good — technical terms are usually defined well, and the narrative is coherent with only small lapses.
8	Very good — clear, precise, and internally consistent; readers can follow the logic with little effort.
9	Excellent — highly precise exposition with strong definitional consistency and near-zero contradictions.
10	Outstanding — crystal-clear, fully consistent, and immediately understandable; every section reinforces a unified narrative.

Table 31: Detailed scoring rubric for **Content Depth**.

Score	Level Descriptor
Critical Analysis — depth of evaluation, fairness of critique, and the comparative assessment of strengths/weaknesses across works.	
1	No critique — purely descriptive with no critical evaluation of methods or findings.
2	Very weak critique — token criticisms that are superficial or inaccurate.
3	Shallow critique — limited or one-sided criticisms lacking evidence or depth.
4	Some critique — occasional meaningful evaluation but inconsistent and incomplete.
5	Moderate critique — reasonable evaluations but lacking systematic comparative rigor.
6	Good critique — fair assessments, identifies key weaknesses and trade-offs in several areas.
7	Strong critique — systematic comparisons, balanced judgments, and evidence-backed critiques.
8	Very strong critique — deep analysis of methodological limitations, assumptions, and empirical gaps.
9	Excellent critique — rigorous, balanced, and insightful comparative evaluation across major works.
10	Masterful critique — transformative critique that clarifies fundamental limits, suggests remedies, and reshapes thinking.
Novelty and Insights — the degree to which the survey produces original synthesis, useful conceptual reframing, or fresh research questions.	
1	No novelty — restates existing content without any new perspective or insight.
2	Very little novelty — trivial observations only; no original framing or notable insight.
3	Low novelty — occasional small observations, but no meaningful new contribution.
4	Limited novelty — a few modest insights but mostly derivative.
5	Some novelty — useful observations that add modest clarity or re-organization.
6	Notable novelty — several original observations or useful reframings emerge.
7	Strong novelty — clear new insights or a helpful conceptual framing that aids understanding.
8	Very strong novelty — original synthesis that changes how parts of the field are seen.
9	Excellent novelty — significant conceptual contribution or new taxonomy with wide applicability.
10	Exceptional novelty — groundbreaking synthesis or insight that opens new directions or major re-interpretation.
Specificity — the granularity and concreteness of the survey's analysis, including whether it goes beyond high-level summaries to discuss concrete implementation details, code-level mechanisms, module interactions, design choices, operational steps, and fine-grained distinctions between methods.	
1	Extremely vague — entirely generic statements with no concrete mechanisms, examples, or fine-grained analysis.
2	Very low specificity — mostly broad labels or superficial lists; almost no detail beyond high-level descriptions.
3	Low specificity — a few concrete terms appear, but the survey remains mostly abstract and non-operational.
4	Below-average specificity — some partial detail is given, but analysis is still dominated by general summaries.

Continued on next page

Table 31 – *Continued from previous page*

Score	Level Descriptor
5	Moderate specificity — the survey includes some concrete mechanisms or examples, but many sections remain at a high level.
6	Fair specificity — several parts discuss implementation or structural details, though the analysis is not consistently fine-grained.
7	Good specificity — the survey regularly addresses concrete methods, internal mechanisms, or module-level distinctions.
8	Very good specificity — detailed analysis of implementation choices, workflow steps, architectural components, or code-level behavior in multiple places.
9	Excellent specificity — highly fine-grained discussion with precise breakdowns of mechanisms, modules, design trade-offs, and operational details.
10	Outstanding specificity — exceptionally granular analysis that closely examines implementation logic, code mechanisms, and detailed system behavior throughout the survey.
Future Directions — the extent to which the survey provides concrete, well-justified, and actionable guidance for future research, including clearly defined problems, feasible methodologies, experimental setups, and evaluation strategies.	
1	Absent — no meaningful future directions or only trivial/obvious statements.
2	Very weak — vague suggestions with no actionable detail or justification.
3	Weak — general directions mentioned but disconnected from identified gaps and lacking feasibility discussion.
4	Limited — some relevant directions, but still broad, under-specified, and difficult to operationalize.
5	Moderate — plausible directions linked to gaps, but limited detail on how to implement or evaluate them.
6	Good — directions are relevant and partially actionable, with some discussion of methods or evaluation.
7	Strong — clearly defined problems with actionable research paths, including hints of methodology or experimental design.
8	Very strong — specific and well-motivated directions, with concrete proposals (e.g., datasets, benchmarks, model designs) and feasibility considerations.
9	Excellent — detailed and insightful research agenda, including prioritized problems, methodological guidance.
10	Outstanding — highly actionable and deeply analyzed roadmap, offering concrete research programs with justified feasibility, step-by-step strategies, and measurable success criteria.

Table 32: Detailed scoring rubric for the **supplementary code evaluation** dimensions used in the CS-domain analysis (Section 3.5).

Score	Level Descriptor
Implementation Fidelity — evaluates whether the survey goes beyond high-level descriptions from papers and captures how methods are realized in practice, including code-level mechanisms, system architecture, module interactions, algorithm implementation, and engineering design choices.	
1	No implementation understanding — the survey only summarizes conceptual ideas from papers without discussing how methods are actually implemented.
2	Very limited implementation understanding — mentions implementation-related terms or repositories but provides no meaningful technical details.
3	Weak implementation understanding — provides superficial descriptions of frameworks, libraries, or general workflows without explaining internal mechanisms.
4	Basic implementation understanding — discusses some implementation aspects but lacks detailed analysis of architectures, modules, or design choices.
5	Moderate implementation understanding — includes several implementation details, but the analysis remains incomplete and mostly descriptive.
6	Good implementation understanding — explains important implementation components, workflows, or algorithm realization for multiple approaches.
7	Strong implementation understanding — provides detailed analysis of implementation mechanisms, module interactions, and practical design decisions.
8	Very strong implementation understanding — systematically connects paper-level ideas with code-level evidence and reveals implementation patterns across methods.

Continued on next page

Table 32 – *Continued from previous page*

Score	Level Descriptor
9	Excellent implementation understanding — provides deep implementation-oriented analysis, including architecture decomposition, design trade-offs, and practical limitations.
10	Outstanding implementation understanding — reconstructs how methods work in practice through comprehensive code-grounded analysis and derives insights unavailable from papers alone.
Practical Reproducibility — evaluates whether the survey provides actionable information required to understand, reproduce, or implement reviewed methods, including code availability, dependencies, experimental configurations, execution procedures, and practical implementation challenges.	
1	No reproducibility support — the survey only describes theoretical concepts without any practical information needed for reproduction.
2	Very limited reproducibility support — briefly mentions implementations or experiments but provides no useful reproduction details.
3	Weak reproducibility support — provides basic information such as datasets or frameworks but lacks practical guidance.
4	Basic reproducibility support — includes some experimental details but misses important requirements such as configurations, dependencies, or implementation procedures.
5	Moderate reproducibility support — provides partial reproduction information, but substantial effort is still required to reproduce the methods.
6	Good reproducibility support — covers important resources, settings, and procedures for reproducing several approaches.
7	Strong reproducibility support — provides detailed information about code resources, implementation requirements, workflows, and practical considerations.
8	Very strong reproducibility support — enables researchers to understand most reproduction requirements and identify potential implementation obstacles.
9	Excellent reproducibility support — provides comprehensive reproduction guidance, including dependencies, configurations, workflows, and limitations.
10	Outstanding reproducibility support — serves as a practical reproduction roadmap that substantially reduces the effort required to implement and validate surveyed methods.

Table 28: Cluster overview for the "AutoSurvey" topic.

Cluster ID	# Papers	Primary Focus (First 30 words)
1	25	*Several papers propose explicit hierarchical structures (e.g., knowledge trees, citation graphs) to organize surveys, while others rely on iterative outline refinement without such explicit models. How do these fundamentally different...*
2	27	*Several papers treat automated survey generation as a one-shot task (generating a complete survey from scratch), while one paper reframes it as a continuous maintenance problem where surveys are incrementally...*
3	33	*Multiple papers identify a fundamental trade-off between structural organization and content depth in automatic survey generation. SurveyLens finds ASG systems excel at structural outlines while Deep Research agents dominate content...*
4	23	*Several papers propose explicit hierarchical structures (e.g., knowledge trees, taxonomies, catalogues) as a foundation for survey generation, while others rely on iterative outline refinement or graph-based planning. What are the...*
5	18	*Several papers propose hierarchical or structured approaches to organizing survey content (e.g., knowledge trees, outlines, attribute trees). How do the design philosophies of these structures differ...*
6	14	*Both <RhinoInsight> and <WebWeaver> propose frameworks to improve deep research agents by introducing explicit control mechanisms over the research process. However, their architectural philosophies differ: RhinoInsight injects control modules...*
7	17	*How do the iterative and recurrent planning mechanisms in <RecurrentGPT>, <IterSurvey>, <SurveyGen-I>, and <AgentCPM-Report> differ in their handling of memory, state updates, and ...*
8	13	*How do the different approaches to structuring and retrieving literature—such as SciPIP's quintuple-based summarization and multi-granularity retrieval versus PaperRobot's knowledge graph construction and link prediction—reflect...*
9	11	*How does the positional bias identified in <Lost in the Middle: How Language Models Use Long Contexts> impact the effectiveness of the natural language memory states used in <RecurrentGPT: Interactive...*
10	15	*How do the retrieval mechanisms in <Paper Circle>, <Language Models Don't Know What You Want: Evaluating Personalization in Deep Research Needs Real Users>, and <SURVEYAGENT> differ in their integration of...*
11	5	*The framework relies heavily on GPT-4 for both instance creation and verification. How might this create a closed-loop system where the evaluation benchmarks are inherently limited by GPT-4's own capabilities...*
12	29	*How do the evaluation philosophies of <SurveyLens>, <SGSimEval>, <SurGE>, <SurveyBench>, and <DeepSurvey-Bench> differ in their definition of a 'high-quality' survey...*